Exhibit 65

EPIDEMIOLOGY

Concepts and Methods

William A. Oleckno
Northern Illinois University

WAVELAND PRESS, INC.
Long Grove, Illinois

CHAPTER SIX

Comparing Measures of Occurrence in Epidemiology

This chapter describes basic procedures for rate adjustment, commonly used measures of association in epidemiology, and the fundamentals underlying statistical significance testing.

Learning Objectives

- Compare and contrast crude, specific, and adjusted rates.
- Explain the rationale for rate adjustment or standardization.
- Perform and interpret the results of age adjustment using both the direct and indirect methods.
- Calculate and interpret each of the following measures of association: risk ratio (cumulative incidence ratio), rate ratio (incidence density ratio), odds ratio, prevalence ratio, percent relative effect, risk difference (cumulative incidence difference), rate difference (incidence density difference), prevalence difference, population rate difference, attributable fraction among the exposed, and population attributable fraction.
- Explain the usefulness of measures of association based on relative and absolute comparisons, respectively.
- Describe the process and rationale for hypothesis testing and determining the statistical significance of an association.
- Define 2000 U.S. standard million population and 2000 U.S. standard population; *a posteriori* comparison; absolute risk; alpha level; clinical (or practical) significance; defined population; directional and non-directional hypothesis; disease odds ratio; excess risk, rate, and prevalence; indirectly standardized rate; population risk difference and population prevalence difference; p-value; referent; relative odds; and relative risk.

of the association. Furthermore, it can tell us whether or not an apparent association is statistically significant, although this is not a reason it is preferred over statistical significance testing. If a confidence interval for a ratio measure of association contains 1.0, then the association is not statistically significant. If for a difference measure of association it contains 0.0, then it is not statistically significant. In short, a confidence interval for a measure of association can be used to get an idea of the probable magnitude and range of the association in the sampled population. If desired, it can also be used to determine whether or not the association is statistically significant, although those who dislike statistical significance testing will not see this as an advantage.

The use of confidence intervals over significance testing has gained many proponents in epidemiology, and confidence intervals are increasingly the preferred way of reporting results in many epidemiologic, public health, and biomedical journals. Nevertheless, statistical testing is still used and has its adherents. Therefore, it is prudent to be familiar with both methods. More is said about significance testing and confidence intervals in chapter 8 and succeeding chapters.*

SUMMARY

- Crude rates are overall, summary measures of occurrence for defined populations. Specific rates are measures of occurrence for distinct subgroups within a defined population, such as age-specific or sex-specific rates. Crude rates are more convenient to compare between populations than specific rates, especially if there are many specific rates to compare. Unlike specific rates, however, crude rates can be distorted (confounded) by differences in the underlying distributions of the populations being compared, particularly age distributions.
- Adjusted rates, like crude rates, are also overall measures of occurrence, but they have been statistically modified (adjusted) to remove the potential distorting effects of one or more factors like age, sex, or race/ethnicity differences between the populations being compared. Adjusted rates permit fair, unbiased comparisons between overall rates.
- There are two basic methods of rate adjustment—the direct and indirect methods. The direct method uses the specific rates in the populations being compared to develop adjusted rates based on a standard population. The indirect method applies the specific rates in the reference population to the study population to develop a standardized mortality (or morbidity) ratio.

*Generally speaking, modern theory in epidemiology holds that study populations do not need to be considered representative samples of some larger general population to provide valid and reliable findings, except in the case of certain cross-sectional studies. This conception, however, can make it difficult to explain the application of confidence intervals and significance testing, which are firmly based on sampling theory. Tools like confidence intervals or tests of significance are used in most epidemiologic studies as relative indicators of the reliability of the findings even if in a technical sense the assumption of random sampling from a normal population is not met.

CHAPTER SEVEN

Association and Causation in Epidemiology

This chapter discusses differences among spurious, noncausal, and causal associations, the various types of causes, and common guidelines used in assessing causation in epidemiologic studies.

Learning Objectives

- Describe and give examples of spurious, noncausal, and causal associations in epidemiology.
- State the common reasons for spurious and noncausal associations, respectively.
- Distinguish among necessary, sufficient, necessary and sufficient, necessary but not sufficient, not necessary but sufficient, and not necessary and not sufficient causes and give examples of each type.
- Describe and give examples of direct and indirect causal associations.
- Briefly describe the causal pie model.
- Discuss six guidelines based on Hill's postulates for judging potential causal associations, including the advantages and limitations of each criterion, respectively.
- Explain the importance of finding causal associations in epidemiology.
- Define predisposing or enabling factors, statistical association, and threshold.

INTRODUCTION

As indicated in chapter 1, one of the primary goals of epidemiology is to discover the *causes** of morbidity and mortality in human populations. This goal has immense practical significance for health professionals because a better

*There are many terms relating to or derived from the root term *cause*. These include causation, causality, causal, causative, cause-effect, etiology, and so forth. These terms are not defined separately in this chapter, but each refers to something similar.

understanding of the causes of morbidity and mortality often leads to more effective prevention, treatment, and control measures and consequently to a reduction in disease incidence, prevalence, or severity.

A *statistical association* between a given exposure and outcome is the starting point for consideration of a causal relationship in epidemiology. A **statistical association** implies that the exposure is related to a change in the *probability* of the outcome. It does not automatically mean that the exposure *causes* the outcome.[1] Hence, a frequently cited maxim in introductory statistics courses is: "Association does not necessarily imply causation." In short, statistical associations should not be accepted at face value. They should be examined for alternate explanations before any conclusions are drawn. Even a statistically significant association (chapter 6) does not guarantee that a true association exists, much less that the association is causal. A **causal association** between an exposure and outcome means that a change in the frequency of the exposure in a population *will result* in a change in the frequency of the outcome, even though not every individual with the exposure will change. A statistical association only implies that those with the exposure are more or less likely to develop the outcome.

To summarize briefly, a valid statistical association means it is *more or less likely* that the outcome will occur in the presence of the exposure, while a valid causal association means that changes in the frequency of exposure will result in changes in the frequency of the outcome. It should be noted that a causal association may be positive (the exposure increases the outcome) or negative (the exposure decreases the outcome). In the former case, the exposure is *hazardous*; in the latter case it is *protective*. The remainder of this chapter focuses on examining statistical associations to determine whether or not they are likely to represent causal associations. Many factors must be considered, and any conclusions must be based on an overall assessment of the evidence.

Types of Association

Statistical associations found in epidemiologic studies (e.g., OR = 3.4) can be categorized into three types. These categories are mutually exclusive.

- Spurious associations
- Noncausal associations
- Causal associations

Spurious Associations

Spurious associations are literally *false* associations. Though they may be found in a particular study population, they are probably due to other explanations. Spurious associations usually result from *random error* (chance) or *bias*, which are discussed more fully in chapter 8. For example, as mentioned in chapter 6, an association is generally considered statistically significant if $p \leq 0.05$. This implies that, assuming there is no association, chance is an

unlikely explanation for the finding given the sample size and strength of the association. Nonetheless, we would still predict that as many as five times out of 100 the association could be due to chance alone. Thus, even statistically significant associations that result from well-executed epidemiologic studies can sometimes be spurious. Inderjit S. Thind, for instance, conducted an ecological study of the association between dietary intake and cancer using a sample of 60 countries. He found a number of significant statistical associations, including some that were biologically implausible and which he thought to be spurious. In his discussion of the findings, he reiterated a common concern in broad-based studies where large numbers of statistical tests of significance are performed. Specifically, he cautioned the readers by stating, "The . . . large numbers of correlations . . . with [some] significant associations occurring purely by chance, suggest extreme care in assessing the role of specific dietary items as risk factors and using the results as the basis for public policy."[21(p162)]

Spurious associations may also arise from sources of bias. *Bias*, which is discussed in chapter 8, is a type of systematic (nonrandom) error in the design, conduct, or analysis of epidemiologic studies, such as the use of flawed measurement techniques, differential recall among study and comparison groups, or selection of study and comparison groups that are dissimilar. Bias can be quite insidious. Consider a hypothetical case-control study of the relationship between exposure to low-frequency electromagnetic fields, such as those generated by electric power lines, electric blankets, and electric alarm clocks, and the incidence of childhood leukemia. The cases consist of patients from area hospitals newly diagnosed with childhood leukemia, and the controls are those without leukemia of similar age, sex, and racial/ethnic background who have been randomly selected from the communities served by the hospitals. The parents of cases and controls are then queried about their children's exposure to low-frequency electromagnetic fields. The parents of the cases may be more likely to recall their children's exposures than those of the controls since they are probably more motivated to remember past exposures that might help explain their children's leukemia than are the parents of the controls. If this is true, the study could result in a *spurious* association between exposure to low-frequency electromagnetic fields and the incidence of childhood leukemia.

Noncausal Associations

Noncausal associations are real associations, but they are *not* causal associations. That is, a change in the frequency of the exposure in a population does not necessarily result in a change in the frequency of the outcome. Noncausal associations often result from *confounding*, which is discussed in chapter 8. The association exists because the exposure is associated with another factor that in turn is associated with the outcome. A whimsical example is provided by Max Michael III, W. Thomas Boyce, and Allen J. Wilcox.[3] Dr. Al Betzerov conducted a prospective cohort study to test his hypothesis that gambling

causes cancer. He chose two neighboring states, one where gambling was legal and the other where it was not. He then followed randomly selected samples of subjects from each state matched by age, sex, urban/rural differences, and family income for 10 years. At the conclusion of the study, he noted a statistically significant positive association between gambling and cancer. Specifically, the residents of Nevada had a higher rate of cancer than those from Utah. The association, although real, was *not* one of cause-effect. Unfortunately for Dr. Betzerov, one of the states he chose was Utah. Utah is a state composed of a large number of Mormons, who have very different lifestyles from typical Nevada residents, who are not Mormons. The fact that the Mormon Church requires its adherents to abstain from tobacco and alcohol explains this association. The apparent causal association between gambling and cancer was due to confounding by alcohol and tobacco use, which are higher in Nevada than in Utah. In other words, alcohol and tobacco use are associated with gambling and are directly linked to cancer. Therefore, although gambling itself does not cause cancer, its association with causes of cancer produces a noncausal association with cancer. This type of association has also been referred to by some as a "spurious association" in that it can lead to an erroneous conclusion about cause and effect.

Risk markers, which were referred to in chapter 1, represent noncausal associations. Although these associations result from confounding with actual risk factors, they are still real associations that have practical significance in screening for disease.[4] For example, calcification in the coronary arteries is a risk marker for coronary heart disease. It does not cause the disease, but it is associated with an increased risk of its occurrence. Its role in coronary heart disease is therefore properly classified as noncausal. Nevertheless, screening for coronary calcium has become an increasingly popular, though controversial, method of detecting possible presymptomatic heart disease (see chapter 13).

Noncausal associations can also result when the defined exposure is a consequence of the outcome instead of the other way around. Hypertension, for example, may result from kidney disease. Thus, one may find a statistical association between hypertension and kidney disease, but in this example, hypertension could not be considered a cause of kidney disease because the exposure does not *precede* the outcome and therefore cannot alter its frequency. In this example, kidney disease is a cause of hypertension. This type of hypertension is generally referred to as secondary hypertension to differentiate it from primary hypertension, which can cause kidney disease.

Causal Associations

Causal associations are those in which changes in the frequency of the exposure in a population produce a change in the frequency of the outcome. In epidemiology, we cannot prove causal associations because it is impossible to account for all the other factors that might play some role in an association, especially in observational studies where there may be many unrecognized,

and therefore uncontrolled, variables. Well-designed experimental epidemiologic studies can come much closer to establishing causation than observational studies, but even in these studies there may be other influential factors of which the investigator is unaware. Since no two humans beings are exactly alike in their makeup or reactions to external stimuli, one cannot always be assured that even randomized groups of people are perfectly comparable. Even laboratory experiments with mice rely on well-defined strains to minimize intraspecies differences that can invalidate the results of an experiment.

A given association may not be conclusively spurious, noncausal, or causal. This is because random error can never be completely eliminated as a possible reason for an association in an epidemiologic study, although it can be greatly minimized. Similarly, it would be extremely difficult to discount any possibility of bias in a study. The same can be said for possible confounding. Thus, the job of the epidemiologist is to determine which type of association is more likely, and this is not always an easy task.

Since our main concern is identifying causal relationships when they exist, we need some guidance in determining whether an association is likely or not to be a causal one. In practice, the determination of a causal association is based on a careful review and judgment of all relevant information available, and never on the basis of one or two studies alone, especially observational studies. It is somewhat like trying a criminal case where there are no eyewitnesses to the crime. The prosecutor has to rely on circumstantial evidence to convince a jury beyond a reasonable doubt that the defendant is guilty. It was based on a thorough review of major epidemiologic and nonepidemiologic studies that in 1964 the Surgeon General of the U.S. Public Health Service first concluded that cigarette smoking is a cause of lung cancer.[5] Before discussing some of the guidelines used to assess potential causal associations, it should be worthwhile to first examine the concept of causation in more detail. This is the subject of the following section.

Types of Causes

With communicable diseases the concept of causation appears to be relatively straightforward. However, as discussed in chapter 3, this apparent simplicity can be deceiving. Not everyone exposed to *Mycobacterium tuberculosis* (the bacterium implicated in tuberculosis), for example, develops tuberculosis. A number of host and environmental factors must also be considered. Similarly, not everyone exposed to cold germs gets a cold. In fact, the more we learn about causation, the more complex it seems. With many noncommunicable diseases, especially chronic conditions like arthritis, mental illness, Alzheimer's disease, multiple sclerosis, cardiovascular disease, diabetes, and so forth, the causal pathways can be extremely complex. Multifactorial etiology (chapter 2) is the rule rather than the exception for most contemporary health-related problems.

Necessary and Sufficient Causes

To get a better understanding of causation as it is commonly used in epidemiology it is helpful to look at different types of causes.* A **necessary cause** is an exposure that is *required* for a particular outcome to occur. Therefore, it is always associated with the outcome. If the exposure is absent, the outcome cannot occur. A **sufficient cause** is an exposure that by itself will produce a particular outcome, but it may not be the only cause of the outcome. Consequently, the outcome may occur without the exposure if the outcome is also caused by other exposures. These two classifications of causes give rise to four possible combinations,[6] which are shown below in the following 2 × 2 table.

		Necessary	
	Sufficient	Yes	No
	Yes	A	C
	No	B	D

Combination A represents a **necessary and sufficient cause**. This is a cause that is required to produce a particular outcome *and* which is able to cause the outcome by itself. This can be represented by:

$$\text{Exposure X} \rightarrow \text{Outcome Y}$$

where Exposure X is the specified cause, and Outcome Y is the specified outcome.

Necessary and sufficient causes are not very common in the real world. One example of a condition that results from a necessary and sufficient cause is lead poisoning. Exposure to lead is *necessary* to produce lead poisoning, and it is also *sufficient*. The rabies virus might also be considered a necessary and sufficient cause of human rabies. It is *not* essential that a necessary and sufficient cause always produces the outcome. Observations have shown, for example, that not everyone presumably infected with the rabies virus contracts the disease even if they have not been immunized.[7] Nevertheless, anyone who contracts rabies must have the virus (i.e., it is necessary), and no other known cause must be present for the disease to occur (i.e., it is sufficient). It is important to emphasize, however, that as knowledge of disease causation expands, classifications may need to be revised. We may learn in the future, for example,

*The types of causes discussed here and subsequently are assumed to be hazardous rather than protective so as to simplify the discussion.

that some causes thought to be necessary and sufficient would be better classified in another way. At one time many believed that cancer was caused by a single factor, still undiscovered. Today we recognize its multifactorial etiology.

Combination B in the above table represents a **necessary but not sufficient cause**. This is a cause that is required to produce a specified outcome *but* is *not* able to cause the outcome by itself. Other causes are necessary for the outcome to occur. This can be represented by:

Exposure X + Other Causes → Outcome Y

Alcoholism is a disease in which alcohol consumption is a necessary but not sufficient cause of the disease. Alcohol consumption is definitely necessary for alcoholism to develop, but other factors, including genetic, social, behavioral, and environmental factors, also appear to be necessary for the disease to manifest itself.

Combination C represents a **not necessary but sufficient cause**. This is a cause that is *not* required to produce a specified outcome *but* when present is able to cause the outcome by itself. This means that there are other causes of the outcome. A not necessary but sufficient cause may be represented by:

Exposure X → Outcome Y and Exposure Z → Outcome Y

where Exposure Z is some other independent cause of Outcome Y. Ionizing radiation at high doses will cause sterility in men. Heavy exposure to certain pesticides will do the same. In this example, Exposure X is ionizing radiation, Exposure Z is a specific pesticide, and Outcome Y is sterility in men. Thus, sterility in men has more than one cause. Both ionizing radiation and certain pesticides are capable of causing sterility in men (at high doses).

Combination D denotes a **not necessary and not sufficient cause**. This is a cause that is *not* required to produce the specified outcome *and* when present is *not* able to cause the outcome by itself. Hence, there are other causes of the specified outcome. A not necessary and not sufficient cause is known as a **contributory cause**. It can be represented by:

Exposure X + Other Causes → Outcome Y and Exposure Z → Outcome Y

where Exposure Z is another independent cause of Outcome Y. Not necessary and not sufficient causes are very common causes of chronic diseases. For example, a sedentary lifestyle is not necessary and not sufficient to cause coronary heart disease (CHD). It is not required for CHD development, *nor* is it considered sufficient to cause CHD by itself. It is, however, a contributory cause of CHD, and when present with certain other contributory causes, such as high blood cholesterol, family history of heart disease, hypertension, cigarette smoking, and so forth, can lead to the development of CHD. That is, the frequency of CHD will be higher in groups with these factors than in groups without them.

A logical extension of this paradigm is one conceptualized by Kenneth J. Rothman and referred to as the **causal pie model**.[8] One can imagine one or

more intact pies neatly divided into several pieces symbolizing what Rothman calls **component causes**. Each pie represents a *sufficient cause* of a particular disease, and each component cause has an essential part in causing that disease. There may be several sufficient causes (pies) made up of various combinations of some of the same and different component causes for any given disease. Whatever the combination, the component causes work together to cause the disease.[8] The causal pie model may remind one of the information asked for on a death certificate regarding the causes of death (see exhibit 2-1 in chapter 2). In a sense, the immediate, antecedent, and underlying causes of death, as well as other significant conditions, seem to parallel the component causes for a particular death.

As intimated earlier, in epidemiology causation is determined by what occurs in populations or groups of people as opposed to what occurs in any particular individual. We know, for example, based on the Framingham Heart Study that people who live certain lifestyles die more frequently from coronary heart disease than those with healthier lifestyles. From the group data, we can make predictions about individuals based on their lifestyle habits, but we cannot expect that the predictions will always be correct. Everyone seems to know someone, for example, who smoked four packs of cigarettes a day, had high blood pressure, and drank like a fish, but lived until 105. Undoubtedly, this person met an "untimely" death when his bungee cord broke after jumping off a bridge. The exception, however, does not make the rule.

Direct and Indirect Causes

Causal associations can also be classified as direct or indirect. A **direct causal association** (or **direct cause**) can be thought of as representing a causal pathway in which there are *no* intermediate variables, while an **indirect causal association** (or **indirect cause**) involves one or more intervening factors.[9] For example, in a direct causal association, X causes Y, where X is the causative exposure, and Y is the outcome. In an indirect causal association, I causes X, which in turn causes Y. While I is a direct cause of X, it is an *indirect* cause of Y. Since I causes X, and X causes Y, it follows that I causes Y based on the definition of a causal association. A change in the frequency of I in a population will result in a change in the frequency of X, which in turn will result in a change in the frequency of Y. Thus, I can be considered an indirect cause of Y.

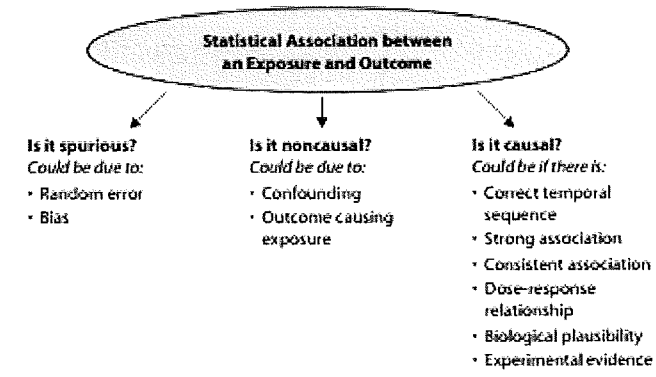
Indirect causes can include a variety of **predisposing or enabling factors** that precede the direct cause. For example, excessive heat applied to the skin is the direct cause of burns, but the exposure to the heat may be influenced by a dangerous working environment or failure to follow certain safety precautions, which might be considered indirect causes of burns. Also, the human immunodeficiency virus (HIV) is said to be the direct cause of AIDS, but factors that facilitate contracting HIV include sharing syringes and promiscuous sexual behaviors. In practice, controlling the predisposing or enabling factors should result in a decrease in frequency of the outcome. Therefore, *predisposing* or *enabling factors* are often referred to as risk factors.

Whatever classification scheme is used, most contemporary health-related problems appear to have multiple causes. This multifactorial etiology, which has been referred to often in this text, presents a challenge to epidemiologists who are concerned with unraveling the determinants of morbidity and premature mortality and to those whose efforts are directed toward their prevention and control. As our knowledge of the natural history of health problems expands, the models of causation and the methods of intervention will continue to undergo change. An interesting article dealing with different conceptions of casusation from an epidemiologic and philosophical perspective is one published in the *Journal of Epidemiology and Community Health* by M. Parascandola and D. L. Weed.[10] While their recommendations may be at odds with many epidemiologists, the discussion itself is can be enlightening, especially for those new to this topic.

Guidelines for Assessing Causation

As shown in figure 7-1, determining whether a statistical association is causal, involves a number of considerations. One must ask if the observed association is likely to be spurious. Random error or bias could explain an association found in a study population. On the other hand, the association could be a noncausal association. Noncausal associations may be due to confounding by an extraneous factor or because the outcome is responsible for the exposure instead of vice versa. Of course, another option is that the association is causal. Okay, you may say, we know the options, but how can we tell if the association is likely to be a causal one? The first step is to examine whether the alternate explanations are plausible. Specifically, is the associa-

Figure 7-1 Deciding Whether an Association Is Likely to Be Causal



tion likely due to random error, bias, confounding, or a reserved causal sequence? This may take some critical thinking, further analysis, or consultation. If these seem to be unlikely explanations, it can be helpful to review some generally accepted guidelines for establishing causation such as those described by Sir Austin Bradford Hill.

In 1965, Sir Austin Bradford Hill, Professor Emeritus of Medical Statistics with the University of London, delivered a landmark address where he outlined nine criteria that could be used to determine if statistical associations were likely to represent causal associations.[31] His reasoning built on the earlier work of others, such as John Stuart Mill, who in 1856 had defined several canons from which causal relationships could be deduced.[6] Over the years many authors have articulated or modified Hill's basic criteria, which have become known as **Hill's postulates**. Using these as a focal point, the following six guidelines should be helpful in deciding whether or not statistical associations are likely to represent causal associations (figure 7-1). In the end, the process of determining causation is largely subjective except for the first guideline, which is actually a requirement.

- **Correct temporal sequence.** In order for an exposure to be considered a cause of an outcome, it must *precede* the outcome. Of all the guidelines used to judge whether an association is causal or not, this is the only one that is considered *absolutely essential*. Exposures that occur concurrently with an outcome or subsequent to an outcome cannot be considered causal because they do not alter the frequency of the outcome. Determining if an exposure precedes an outcome can be problematic in cross-sectional studies where exposure and outcome are assessed concurrently. For example, in a cross-sectional study designed to determine if there is a relationship between the prevalence of excess body weight and osteoarthritis, it may not be clear which factor came first. Thus, the correct temporal sequence cannot be established reliably. This can also be a problem in case-control studies where the prevalence of the outcome is assessed instead of its incidence.
- **Strength of the association.** In general, the stronger an association between a given exposure and outcome (see table 6-3), the more likely the association is causal. When the risk ratio is very high, for example, it is more difficult to explain away the association due to unrecognized or subtle sources of bias or confounding. Compared to nonsmokers, those who smoke and are exposed to high levels of asbestos in their jobs have a fifty- to ninety-fold increased risk of lung cancer. It seems improbable that these factors are not causative. Even if some bias or confounding exists, it is unlikely that it would account for the entire relationship. This is not to say that small associations cannot also be causal in nature. This is one reason why several guidelines are needed to assess causality.
- **Consistency of the association.** When other investigators studying different populations at different times in different places using different methodologies obtain similar findings with regard to a specific association, it

increases the probability that the association is causal. In concluding that cigarette smoking is a cause of lung cancer, the Advisory Committee to the Surgeon General of the United States cited diverse epidemiologic and other studies showing a strong relationship between smoking and lung cancer.[5] One way of determining if an apparent association is likely to be due to random error is to replicate the study. If the findings are consistent, it strengthens the case for a causal association, assuming there are no significant sources of bias or confounding in the studies.

- **Dose-response relationship.** In general, if increased levels of exposure lead to greater frequencies of the outcome, then this is suggestive of a causal relationship. Heavy smokers, for example, have been shown to be at a higher risk of lung cancer than light smokers. In fact, a linear dose-response relationship between smoking and lung cancer can be demonstrated based on the number of cigarettes smoked per day. The absence of a dose-response relationship does not necessarily mean that an association is non-causal, however. A threshold may exist. A **threshold** is a level of exposure (dose) that must be reached before effects become apparent. Below the threshold, there are no observed effects. Copper, which may be found in small quantities in drinking water and certain foods, demonstrates a threshold; that is, copper has no adverse effects until it reaches a certain level in the body. In fact, in very small quantities it is an essential mineral needed for proper growth and development. On the other hand, a dose-response relationship could be due to a strong confounding factor that closely follows an exposure.[12] Once again, several guidelines should be considered in assessing causation.
- **Biological plausibility.** The basic question here is, does the association make biological sense? Is the association credible based on our understanding of the natural history of the disease or possible pathogenic mechanisms? When Thind found significant associations for protein, fat, and caloric intake and certain forms of leukemia, he could offer no biological evidence to support the associations, thereby casting doubt on their authenticity.[2] Failure to make biological sense, however, does not necessarily negate the possibility of a causal association. In some cases, our understanding of the biological mechanisms may be incomplete, and what does not make sense today may make sense sometime in the future. From a contemporary vantage point, it seems difficult to understand why the theory of contagion was considered controversial as an explanation for the spread of epidemics during the Middle Ages.
- **Experimental evidence.** Having experimental evidence to support an association between a given exposure and outcome strengthens the case for a causal association. Well-designed randomized controlled trials, for example, can provide strong corroboration of a suspected causal association. This is because this study design, properly implemented, can virtually eliminate selection bias and confounding as alternate explanations for a causal

association (see chapters 8 and 12). Of course, the degree of control possible in epidemiologic experiments is not to the same level as that in animal studies. Nevertheless, they can be powerful tools for establishing causation. Evidence from nonepidemiologic experiments can also be used in assessing cause-effect relationships. Because of the limited circumstances in which experimental studies can be conducted with humans, some associations will not be testable in this manner. We would not perform a randomized controlled trial on the effects of microwave radiation on cataract development, for example, because such a study would be unethical even if some were willing to volunteer for the investigation.

Table 7-1 ranks the most common types of epidemiologic studies in descending order of the degree to which identical findings of a statistical association are likely to demonstrate a causal association. The ranking is based on the relative probability of encountering unrecognized bias, confounding, or other errors within the specific study designs. It also assumes that the studies have been planned appropriately and conducted to minimize errors. A poorly designed experimental study can provide less convincing evidence of causality than a well-designed observational study. It should be kept in mind, however, that causality is never determined based on the findings of one study alone. Causation is a judgment based on relevant, cumulative information. Meta-analyses (chapter 12) have provided some hope of reaching more definitive conclusions in epidemiologic studies. Whether they will fulfill this hope depends on the care in which they are designed, implemented, and interpreted.

Table 7-1 Ranking of Common Epidemiologic Studies in Terms of the Relative Probability that the Findings Represent Causal Associations

1. Randomized Controlled Trial	5. Case-Control Study
2. Group Randomized Trial	6. Cross-Sectional Study
3. Prospective Cohort Study	7. Ecological Study
4. Retrospective Cohort Study	8. Descriptive Study

SUMMARY

- Statistical associations found between given exposures and outcomes can be of three types—spurious, noncausal, or causal. Spurious associations are false associations that are usually due to random error or bias. Noncausal associations usually result from confounding, although they can also occur when the exposure is the result of the outcome instead of the other way around. Risk markers represent noncausal associations that have practical value in screening for disease. Causal associations are ones in which a change in the frequency of the exposure results in a change in the frequency of the outcome in a population.

- Causes can be classified as to whether or not they are necessary and/or sufficient and whether they are direct or indirect. A necessary cause is one that is required to produce an outcome, while a sufficient cause is one that can produce the outcome by itself (i.e., in the absence of other known causes). The most common types of causes are those that are not necessary and not sufficient. These are known as contributory causes and are the causes that account for most contemporary health-related problems. The causal pie model expands upon the not necessary and not sufficient causes by considering a constellation of component causes that are sufficient to cause disease. Direct causes do not involve any intermediate factors in the causal pathway. Indirect causes include a variety of predisposing or enabling factors that precede the direct cause of an outcome. Controlling indirect causes can reduce the incidence of particular outcomes and is sometimes easier than controlling the direct causes.
- Because it is not possible to prove causation directly, it is helpful to have reliable guidelines upon which to judge a statistical association in terms of its likelihood of being causal. A final decision regarding causation should be based on all relevant information and not just on the basis of one or two studies, especially observational studies. Six guidelines, derived from Hill's postulates, should help in determining whether an association is likely to be causal. These guidelines are correct temporal sequence, strength of the association, consistency of the association, dose-response relationship, biological plausibility, and experimental evidence. Of these guidelines, only correct temporal sequence is required for an association to be considered causal. The others are highly suggestive of causation, however, especially when all or most of them are met.

New Terms

- biological plausibility
- causal association
- causal pie model
- component causes
- consistency of the association
- contributory cause
- correct temporal sequence
- direct causal association
- direct cause
- dose-response relationship
- experimental evidence
- Hill's postulates
- indirect causal association
- indirect cause
- necessary and sufficient cause
- necessary but not sufficient cause
- necessary cause
- noncausal association
- not necessary and not sufficient cause
- not necessary but sufficient cause
- predisposing or enabling factors
- spurious association
- statistical association
- strength of the association
- sufficient cause
- threshold

Study Questions and Exercises

1. For each of the following statements indicate whether the results are more likely to be due to a spurious association, a noncausal association, or a causal association. Also, explain the reasons for your answers.
   a. A case-control study revealed that there was a moderate to strong association between coffee consumption and deaths from coronary heart disease. Other studies have shown that those who drink coffee are more likely to smoke than those who do not drink coffee.
   b. A prospective cohort study showed that women who exercise regularly were less likely to contract cancer than women who exercised only occasionally or not at all. The exercise group was selected from women attending a fitness center, and the comparison group was selected from women attending a weight-loss clinic.
   c. A large randomized controlled trial showed that folic acid supplementation by prospective mothers significantly reduced the incidence of neural tube defects in their offspring. This finding was confirmed in subsequent studies.
   d. A large exploratory epidemiologic study examined the possible relationship of 25 different lifestyle behaviors to teenage suicide. One of the findings was a positive association between bicycle helmet use and suicide ($p = 0.05$) that had not been previously reported in the literature.
2. On bottles of wine and other alcoholic beverages, it states, "According to the Surgeon General, women should not drink alcoholic beverages during pregnancy because of the risk of birth defects." Discuss the evidence that alcohol consumption causes birth defects using the six guidelines for causation discussed in this chapter. For each guideline, describe the degree to which the evidence supports a conclusion of causation and the reasons for your response. In answering this question it may be necessary to consult a review of epidemiologic literature on alcohol consumption and birth defects.
3. Provide an example other than one used in this chapter of a necessary and sufficient cause, a necessary but not sufficient cause, a not necessary but sufficient cause, and a not necessary and not sufficient cause of disease, respectively. Also indicate why your examples are appropriate.
4. Give two examples, respectively, of direct and indirect causes of disease and justify your choices.

References

1. Vogt, W. P. (1999). *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences*, 2nd ed. Thousand Oaks, CA: Sage Publications.
2. Thind, I. S. (1986). Diet and Cancer—An International Study. *International Journal of Epidemiology* 15(2): 160–162.

3. Michael, M. III, Boyce, W. T., and Wilcox, A. J. (1984). *Biomedical Bestiary: An Epidemiologic Guide to Flaws and Fallacies in the Medical Literature*. Boston: Little, Brown, and Company.
4. Szklo, M., and Nieto, F. J. (2000). *Epidemiology: Beyond the Basics*. Gaithersburg, MD: Aspen Publishers, Inc.
5. U.S. Department of Health, Education, and Welfare (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. USPHS Publication No. 1103. Washington, DC: U.S. Government Printing Office.
6. Last, J. M., ed. (2001). *A Dictionary of Epidemiology*, 4th ed. New York: Oxford University Press.
7. Chin, J., ed. (2000). *Control of Communicable Diseases Manual*, 17th ed. Washington, DC: American Public Health Association.
8. Rothman, K. J. (2002). *Epidemiology: An Introduction*. New York: Oxford University Press.
9. Jekel, J. F., Elmore, J. G., and Katz, D. L. (1996). *Epidemiology, Biostatistics, and Preventive Medicine*. Philadelphia, PA: W. B. Saunders Company.
10. Parascandola, M., and Weed, D. L. (2001). Causation in Epidemiology. *Journal of Epidemiology and Community Health* 55: 905-912.
11. Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58: 295–300.
12. Brownson, R. C., Remington, P. L., and Davis, J. R. (1998). *Chronic Disease Epidemiology and Control*, 2nd ed. Washington, DC: American Public Health Association.

CHAPTER EIGHT

Assessing the Accuracy of Epidemiologic Studies

This chapter deals with the accuracy of epidemiologic studies, specifically validity and precision. In particular, threats to accuracy in the forms of bias, confounding, and random error are examined.

Learning Objectives

- Define and explain accuracy, validity, and precision.
- Compare and contrast internal and external validity.
- Distinguish between selection and information bias.
- Identify potential types of selection bias based on study descriptions.
- Identify potential types of information bias based on study descriptions.
- Differentiate between differential and nondifferential misclassification and the potential consequences of each.
- Identify basic methods of controlling selection and information biases, respectively.
- Explain the concept of confounding, the requirements for confounding, and the potential consequences of confounding.
- Identify specific methods to minimize confounding.
- Define random error and its major components.
- Describe the major methods of assessing random error, including their relative strengths and weaknesses.
- Explain two methods of reducing random error in a study.
- Define beta level; error; individual, pair, and frequency matching; interval estimation; systematic and nonsystematic error; positive and negative bias; positive and negative confounding; potential confounder; power; probability sample; residual confounding; Simpson's paradox; source population; and type I and type II errors.

In other words, researchers seeking to understand the causes of homelessness by studying individual risk factors for homelessness are committing a type III error by giving the right answer to the wrong problem.

The Second Definition

Another definition of a type III error has to do with statistical significance or hypothesis testing. Basically, it occurs when one rejects the null hypothesis (H_O), and the alternate hypothesis (H_A) is directional but in the opposite direction of the true association. For example:

H_O: OR = 1.0

H_A: OR > 1.0

If one rejects the null hypothesis that the odds ratio equals one in favor of the alternate hypothesis that the odds ratio is greater than one, but in reality the true odds ratio is less than one, then the researcher has committed a type III error. This concept of a type III error is due to sampling variation (i.e., random error). Theoretically, a larger sample should lead to a lower probability of making a type III error. L. Leventhal and C. Huynh (1996) have indicated that this kind of a type III error can also occur when the alternate hypothesis is non-directional as shown below.

H_O: OR = 1.0

H_A: OR ≠ 1.0

Even though the alternate hypothesis is non-directional, some researchers will assume that the direction of the relationship in their study population is the correct one. If it is not, a type III error has been committed.

References: Kimball, A. W. (1957). Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association* 52 (278): 133–142; Leventhal, L., and Huynh, C. (1996). Directional Decisions for Two-tailed Tests: Power, Error Rates, and Sample Size. *Psychological Methods* 1 (3): 278–292; Schwartz, S., and Carpenter, K. M. (1999). The Right Answer for the Wrong Question: Consequences of Type III Error for Public Health Research. *American Journal of Public Health* 89 (8): 1175–1180.

As implied in chapter 6, statistical significance (hypothesis) testing has a number of limitations and has found disfavor among many epidemiologists. Some common objections are: (a) the cut-off point for determining statistical significance is completely arbitrary; (b) the p-value is confounded by the effects of sample size and strength of the association; and (c) statistical significance is often misinterpreted. With regard to the first point, one can argue convincingly that there is very little difference between p = 0.04 and p = 0.06. Yet, with an arbitrary cut-off point of p = 0.05, two different conclusions would be reached based on these similar values. The second point is important because a p-value will vary simply because of differences in sample size or the strength of the association. For example, a very large sample tends to produce very small p-values, while a very small sample tends to produce large p-values. Thus, even weak associations may be considered statistically significant if the sample size is large enough but not statistically significant if the sample size is smaller. Also, a relatively strong association may be accompa-

Exhibit 8-4
Using P-values to Assess the Potential Effect of Random Error on an Association

Is $p \leq 0.05$?	Is $p > 0.05$?
1. If YES, the association is *statistically significant*. The probability of a type I error is equal to or less than 5%, which is the traditional cut-off point (alpha level).	1. If YES, the association is *not statistically significant*. The probability of a type I error is more than 5%, which is the traditional cut-off point (alpha level).
2. Random error is an *unlikely* explanation for the observed association.	2. Random error is a *reasonable* explanation for the observed association.
3. The *smaller* the p-value, the *less* likely random error explains the association.	3. The *larger* the p-value, the *more* likely random error explains the association.
4. *If* the sample size is *very large*, most associations will be statistically significant, so the association should be judged for its *practical significance*.	4. *If* the sample size is *small*, the association should be considered *inconclusive* due to low power and possible type II error.

Note: This framework assumes that bias and confounding are not responsible for the finding.

nied by a large p-value because of small sample size. In isolation one cannot be sure if a p-value is more a reflection of sample size or strength of the association or both. Finally, "statistically significant" is often misinterpreted as meaning that the null hypothesis is false or that the association is one of cause and effect. Neither can be demonstrated using statistical significance testing, which depends on probabilities. For these and other reasons most epidemiologists prefer using confidence intervals over statistical significance testing when it comes to assessing random error.[16, 24] The basic methods are discussed in the following section.

Assessing Random Error Using Confidence Intervals

Random error can be readily assessed using confidence intervals, which were introduced in chapter 5. This method is referred to as **interval estimation**. Statistically speaking, a confidence interval is constructed around a *point estimate* of the population parameter for a given level of confidence, usually 95 percent. The extent of random error in the estimate is judged by the width of the confidence interval. If the confidence interval is fairly narrow, *and* the confidence level is high, this suggests that there is little random error in the estimate. Therefore, the point estimate can be considered relatively precise. Conversely, if the interval is fairly wide, *and* the confidence level is high, it implies that there is significant random error, and the point estimate can be considered relatively imprecise. Two important caveats need to be kept in mind. First, in assessing the extent of random error it is important to have a

high level of confidence since confidence intervals tend to narrow as the confidence level decreases. It is much easier, for example, to be 50 percent confident that a point estimate falls within a relatively narrow range of values than to be 90 or 95 percent confident. Second, confidence intervals tell us nothing about whether or not a measure of association is valid. Bias or confounding may affect a measure of association no matter how precise it may appear based on a confidence interval.

Assuming there is no significant bias or confounding present, if RR = 2.5, and the 95% CI = 2.1 to 2.9, we would be reasonably confident that we have a relatively precise estimate of the measure of association, which in this case happens to be a risk ratio. On the other hand, if RR = 2.5, and the 95% CI = 0.6 to 19.5, we would say that the measure of association appears very imprecise (i.e., subject to substantial random error). These comparisons depend on using the same confidence level, which is usually, but not always, 95 percent. Confidence intervals are best viewed as general indicators of the amount of variability in a measure.[24] If the effects of bias and confounding have been adequately prevented or controlled, however, and if the confidence level is high enough (e.g., 95 percent), then we can be reasonably sure that a narrow confidence interval means that we probably have a relatively accurate measure of association in our study population. Specific methods for calculating confidence intervals for measures of association are discussed in subsequent chapters dealing with specific study designs.

Confidence Intervals and Significance Testing

Primarily for reasons cited in the previous section, many epidemiologists favor interval estimation over statistical significance testing. Confidence intervals can be used to assess statistical significance, but this is not their attraction, since those who favor confidence intervals do not care to do significance testing. For the record, however, a confidence interval for a risk, rate, odds, or prevalence ratio containing a value of one represents a *non-statistically significant* finding, while a confidence interval *not* containing a value of one represents a *statistically significant* finding. Similarly, a confidence interval for a risk, rate, or prevalence difference containing zero is *not statistically significant*, while one *not* containing zero is *statistically significant*. The value of confidence intervals, however, lies in the fact that they provide information not readily available from significance testing. For one thing, confidence intervals give us a range of possible values for the measure of association. RR = 3.4 (95% CI = 0.7 to 15.1) indicates that while the point estimate of the population RR is 3.4, we are 95% confident that the true population RR ranges from 0.7 to 15.1, assuming no systematic errors are present. Furthermore, as stated earlier, we can see from this broad range of values that the estimate is relatively imprecise, and, therefore, we would not want to stake too much on RR really being 3.4. RR = 2.6 (95% CI = 2.1 to 3.0), on the other hand, suggests a relatively precise estimate, and we would be relatively confident using 2.6 as our estimate of the RR knowing that we have probably not under- or over-

estimated it by too much, assuming again that there are no systematic errors. Of course, on average, there is still a five percent chance that the true value is really outside this range (i.e., we are after all using a 95 percent versus a 100 percent confidence interval, which would be too broad to be useful). Certainty is just not part of the game plan.

Unlike p-values generated from significance testing, confidence intervals provide clues to the magnitude of an association *and* the precision of the point estimate. This information is lost in p-values.[16] Thus, a 95 percent confidence interval of 0.9 to 11.7 tells us much more than a p-value stated as $p > 0.05$ or even $p = 0.07$ when the alpha level is 0.05. With the latter information alone, we would be forced to conclude that the finding is not consistent with the null hypothesis without knowing whether the inconsistency is due to small sample size or a weak or nonexistent association. With the former information, however, we could speculate that the association is probably positive and moderate to strong based on the propensity toward high, positive values in the confidence interval. We could also speculate that the sample size that produced the confidence interval is somewhat small given that the interval is relatively wide. The width of a confidence interval is proportional to sample size, which in turn is proportional to the level of precision, all other factors being equal. Thus, the stated confidence interval appears to reflect significant random error and low precision in the estimate based on its width. Of course, like the selection of the alpha level in statistical significance testing, selection of the confidence level is arbitrary. Also, as with statistical significance testing, interval estimation may be influenced by sources of systematic error, and neither is sufficient for establishing causation between an exposure and outcome. Because of the potential for systematic errors, some recommend that bias and confounding be addressed in a study prior to assessing random error.[21]

In conclusion, confidence intervals, and to a lesser extent p-values, can be used to assess random error in study findings. However, because of the potential for uncontrolled systematic errors and for reasons related to the validity of assumptions regarding the statistical model being employed, these methods are best used to make *qualitative* versus *quantitative* decisions about the *relative* amount of random error.[24] Of these two options, confidence intervals have clear advantages over statistical significance (hypothesis) testing. Why then does hypothesis testing persist? Marks R. Nester[25] has suggested some possible explanations: (a) the appearance of objectivity and exactitude; (b) the availability of easy-to-use statistical software; (c) traditional teaching practices; and (d) demands of certain journal editors or thesis directors. Additionally, he seemed to imply that "peer pressure" may be a factor when he included an explanation that "everyone else seems to use them [tests of hypotheses]."[25(p401)]

Reducing Random Error

There are two major methods of reducing random error in a study, one of which has already been mentioned.

Exhibit 66

Introduction to Meta-Analysis

Michael Borenstein
Biostat, Inc, New Jersey, USA.

Larry V. Hedges
Northwestern University, Evanston, USA.

Julian P.T. Higgins
MRC, Cambridge, UK.

Hannah R. Rothstein
Baruch College, New York, USA.

WILEY
A John Wiley and Sons, Ltd., Publication

Registered office
W & S sL T A S s W s Suss x P 9 8SQ g

L bra f C ngress Catal gu ng n Pub cat n Data

s s / M B s
p ;
s g p s x
SBN 978 0 470 05724 7
M s s B s M
DNLM M A s s s T p WA 950 6 4 2009
853 M48 58 2009
6 0 72 22

2008043732

A g s k s v B s L

SBN 978 0 470 05724 7 B

S 0 5/ 3p T s g S w S v s Pv L P
P P p L 4YY

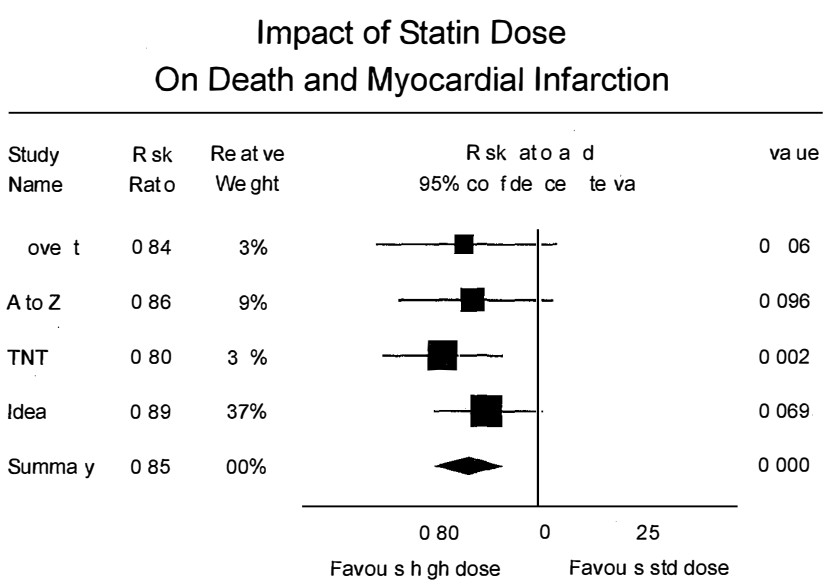


Figure 1.1 High-dose ve sus s anda d-dose of s a ins (adap ed f om Cannon *et al.*, 2006).

w rk w e e ec s zes ssess e c s s e cy e e ec cr ss s d es d c p e s ry e ec

T e e ec s ze c ld represe e p c erve s c s e p c ed c l re e sk ec e p c e c g e d es sc res r e p c ew pr c l e ber s l s ccess lly re g ps re T e e ec s ze s l ed e p c erve s b c ld represe *any r lationship* be wee w v bles s c s ed ere ce es sc res r les vers s les e d ere ce c cer r es r pers s exp sed r exp sed sec d d s ke r e d ere ce c rd c eve s r pers s w w d s c pers l y ypes I c w we ge er lly c ll *:ffi ct siz* c ld re er s ply e es e s gle v l e s c s e prev le ce Ly e d se se

I s ex ple e e ec s ze s e sk r A r sk r 1 0 w ld e e sk de r MI w s es e b gr ps w le r sk r less 1 0 w ld e er sk w s l wer e g d se gr p d r sk r gre er 1 0 w ld e er sk w s l wer e s d rd d se gr p

T e e ec s ze r e c s dy s represe ed by sq re w e l c e sq re represe g b e d rec d g de e e ec Here e e ec s ze r e c s dy lls e le ce er (d c g be e r e g d se gr p) T e e ec s s r ges (s d s e ce er) e *TNT* s dy d we kes e *Id al* s dy

N e F r e s res e ec s ze b sed r s (s s ex ple) r 1 0 represe s d ere ce be wee gr ps F r e s res e ec b sed d ere ces (s c s e d ere ce) d ere ce 0 0 represe s d ere ce be wee gr ps

e ec size e ig f cen e in ic es c n l ien s e e e likely s vive
T e l se ves ig lig e ll ing in s

T e e ec sizes e e s n bly c nsis en s y s y M s ll in e nge f 5 9 , ic s gges s i l be i e c e s y e ec size
T e s y e ec is isk i f 79 i 95% c n ence in e v l f 72 87 (is, 2 % ec e se in sk f e , i 95% c n ence in e v l f 3% 28%) T e v l e f es y e ec is 8
T e c n ence in e v l b n s e c e ec size in ic es e ecisi n in s y If e in e v l excl es , e v l e is less n 5 n es y is s is ic lly signi c n Six f e s ies e e s is ic lly signi c n ile 27 e e n

In s , e e en e ces e isk f e by s e 2 % An , is e ec s e s n bly c nsis en c ss ll s ies in e n lysis
Ove e c se f is v l e e ex l in e s is ic l ce es le ese c ncl si ns O g l in e esen c e is si ly ex l in e n lysis es e ese ec nis s, e e s e n ive evie es n T e key i e ences e s f ll s

STAT ST CAL S GN F CANCE

One f e s q es i ns ske f s y is e s is ic l signi c nce f e es l s T e n ive evie s n ec nis f syn esizing e v l es e i e en s ies, n s e l i e s isc e e ieces f In is ex le six f e s ies e e s is ic lly signi c n ile e e 27 e e n , ic le s e c ncl e e e s evi ence g ins n e ec , e es l s e e inc nsis en (see v e c n ing in C e 28) By c n s, e e n lysis ll s s c bine e e ec s n ev l e e s is ic l signi c nce f e s y e ec T e v l e es y e ec is $p =$ 8
W ile ne ig ss e 27 s ies f ile e c s is ic l signi c nce bec se ey e e s ll e ec s, i is cle e f es l is is n e c se In f c , e e en e ec in ny f ese s ies s c lly *larger* n e e en e ec in e six s ies *were* s is ic lly signi c n R e , e e s n 82% f e s ies e e n s is ic lly signi c n is ese s ies s ll s le sizes n l s is ic l e In f c , s isc sse in C e 29, s e f less n 2 % By c n s , e e e n lysis excee e 99 9% (see C e 29)
As in is ex le, if e g l f syn esis is es e n ll y esis, en e n lysis vi es e ic lly ig s ec nis is se H eve , e n lysis ls ll s s ve bey n e q es i n f

s s s n n ss qu s ns h n s n n s
v n

CL N CAL MPORTANCE OF THE EFFECT

S n h n u n v v w s usu h v u s
h v us s u s h v w w n us n h qu s n wh h n
h v n ws us j h nu h h s s Th s n
h n s s uss n h n u h B n s h
n h s s uss n h s v u w us u n s
s z h s u n h s s z s h h n s s
Th s s n us h s z s wh w u I n n
n n s s n u wh h n n h
w n n w h n u s h s h 5% % 2 % n
h s s h n n h s z S w h n n
n n n n v n n s h s s s s u n s u h
nu n ns n s juv n s n s h su v v
n s w h n n h qu s n w s s u h n u
h Th v u n us n h h s n z n s
h h s n z s ss h n

CONS STENCY OF EFFECTS

Wh n w w n w h n s u s s n
s wh h n h s z s n s s n ss s u s Th ns
qu n u h n s s n u s h s h 2 %
s w h u h u s h s h 2 % n v u
h n s s h s 2 % n s u ns wh u n 6 % n
h s
Th n v v wh s n h n s ss ss n h n s s n
s Th n v v w s s w h v u s n us h v u s v n
h s z s u s w s h n h s u h h n s u
v u n n h v u 5 s n n
h h w s n Th v u *could*
s z u u s s n s u (s h
GISSI s u n F u 2 x) Th v u 5 *could* s
(n) s z u u s n s s u (s h
F h s u x)
Th s n s n ss n n v v w s O n s h s n
n n s n n s u n h s n I s s u s s s
s n n wh h s n h v w s s h s u s s n n Th s
u n s h u h n s s h T w h s C
G n's h n *Arsenic and Old Lace*, w h s h s

e esul s e e e es ul ssu e c e ec s ze T e e e these cases the random ealects model is more easily justialed than the alxed-ealect el

A l e g l s l ss s usu ll ge e l ze ge sce s T e e e e ke e gu e ll e s u es use e c l l e e p pul e ul be p ss ble ex p l e sp pul es eu l e l s s ul be seve el l e

A caveat

T e e s e c ve e b ve I e u be su es s ve s ll e e es e e be ee s u es v ce (τ^2) ll ve p p ec s W le e e ec s el ss ll e pp p e el e l ck e ee e ppl c ec l I s c se e ev e e c se g seve l p s e c e p ble c

O e p s ep e sep e e ec s *not* ep su e ec T e pe s e e e ll u e s e c c clus s b u ee ec s ze sc e ce e v l T e p ble s s e e es ll eve v e c u g (see C p e 28) p ss bl e c e e us c clus

A e p s pe xe e ec l s s T s pp c ul el esc p ve l s s e clu e s u es bu ul ll us ke e ces b u e p pul T e p ble s pp c s ( ) e ke e e ces b u e p pul (b) e e s ll ke ese e e ces eve e e e

A p s ke B es pp c e e ees e τ^2 sb se u s e ecu e se s u es T s sp b bl e bes p bu e p ble s el vel ese c e s ve expe se B es e l s s A ll s e ese c e s ve p l s p c l b ec s pp c

F e ge e l scuss s ssue see *When does t make sense to pe form a meta-analys s* C p e 40

MODEL SHOULD NOT BE BASED ON THE TEST FOR HETEROGENEITY

I e ex c p e e ll uce es e ull p es s e be ee s u es v ce s ze T s es s b se e u be ee s u es v ce bse ve el ve e u e ul expec e s u es c u ll s e c e ec s ze

S e ve p e e p c ce s g xe e ec el e s c g e ec s el e es ge e s s s c ll s g c T s p c ce s ul be s gl sc u ge bec use e ec s use e e ec s el s ul be b se u u e s g e e ll s u e s s e c e ec s ze e u c e s s c l es (espec ll s ce e es e e ge e e su e s l p e)

CHAPTER 28

Vote Counting – A New Name for an Old Problem

Introduction
Why vote counting is wrong
Vote counting is a pervasive problem

INTRODUCTION

One question we often ask of the data is whether or not it allows us to reject the null hypothesis of no effect. Researchers who address this question using a narrative review need to synthesize the *p*-values reported by the separate studies. Since these are discrete pieces of information and the narrative review provides no statistical mechanism for synthesizing these values, narrative reviewers often resort to a process called vote counting. Under this process the reviewer counts the number of statistically significant studies and compares this with the number of statistically nonsignificant studies.

In some cases this process has been formalized, such that one actually counts the number of significant and nonsignificant *p*-values and picks the winner. In some variants, the reviewer would look for a clear majority rather than a simple majority. Or, the reviewer might not work directly with the *p*-values, but with the discussion section of the papers which are based on the *p*-values.

One might think that summarizing *p*-values through a vote-counting procedure would yield more accurate decision than any one of the single significance tests being summarized. This is not generally the case, however. In fact, Hedges and Olkin (1980) showed that the power of vote-counting considered as a statistical decision procedure can not only be lower than that of the studies on which it is based, the power of vote counting can tend toward zero as the number of studies increases. In other words, vote counting is not only misleading, it tends to be *more* misleading as the amount of evidence (the number of studies) increases!

Introduction to Meta-Analysis M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein

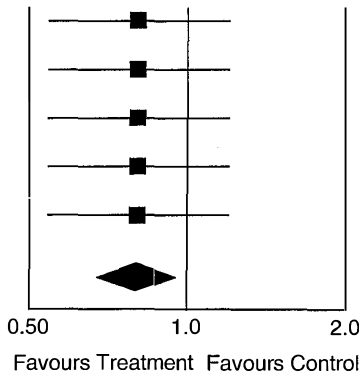
I , g w
q w (g ,
w w g
) O g x w
, w x

WHY VOTE COUNTING IS WRONG

g g g g
x , w g g W
, W g g *could*
, w w
P ,
z z E , w
g z q I w ,
, the absence o a stat st cally s gn cant effect s not ev dence that an effect s absent.

F x , z (RC)
, w g (
65) F g 8 1 5
g , g
B , (F g 8 1), g
g , w N
, w
W w g , z
, w g
65,

	RR	*N*	*p*-value	Risk ratio and 95% confidence interval
RCT A	0.80	800	0.265	
RCT B	0.80	800	0.265	
RCT C	0.80	800	0.265	
RCT D	0.80	800	0.265	
RCT E	0.80	800	0.265	
Summary	0.80	4,000	0.013	



Figu e 28.1 > <

0.013. Clearly, the absence of significance in each study is due to a lack of precision rather than a small effect.

For purposes of explaining why vote counting is a bad idea, we could end the chapter here. However, because vote counting in its various forms is so pervasive, we will expand on this idea to show how the basic mistake that underlies vote counting affects much of the literature, and how meta-analysis can help address this problem.

VOTE COUNTING IS A PERVASIVE PROBLEM

While the term vote counting is associated with narrative reviews it can also be applied to the single study, where a significant *p*-value is taken as evidence that an effect exists, and a nonsignificant *p*-value is taken as evidence that an effect does not exist. Numerous surveys in a wide variety of substantive fields have repeatedly documented the ubiquitous nature of this mistake.

In medicine, for example, Freiman, Chalmers, Smith and Kuebler (1978) surveyed reports of controlled clinical trials that had been published in a number of medical journals (primarily *The Lancet*, the *New England Journal of Medicine*, and the *Journal of the American Medical Association* during the period 1960–1977), and selected 71 that had reported negative results. The authors found that if the true drug effect had been in the region of 50% (e.g. a mortality rate of 30% for placebo vs. 15% for drug), median power would have been 60%. In other words, even if the drug cut the mortality rate in half there was still a 40% probability that the study would have failed to obtain a statistically significant result.

The authors went on to make the following point: Despite the fact that power was terribly low, in most cases the absence of statistical significance was interpreted as meaning that the drug *was not effective*. They wrote: 'The conclusion is inescapable that many of the therapies discarded as ineffective after inconclusive "negative" trials may still have a clinically meaningful effect' (p. 694). In fact, it is possible (or likely) that some of the therapies discarded on this basis might well have had very substantial therapeutic effects.

In the social sciences Cohen (1962) surveyed papers published in the *Journal of Abnormal and Social Psychology* in 1960. Mean power to detect a small, medium, or large effect, respectively, was 0.18, 0.48, and 0.83. Cohen noted that despite the low power, when the studies with *negative* results are published, readers tend to interpret the absence of statistical significance as evidence that the treatment has been proven ineffective.

In the years that followed a kind of cottage industry developed of publishing papers that documented the fact of low power in any number of journals in the area of behavioral research. Many of these are cited in Sedlmeier and Gigerenzer (1989) and Rossi (1990). Similar papers were published to document the same problem in the field of medicine (Borenstein, 1994; Hartung, Cottrell & Giffen, 1983; Phillips,

Scott, & Blasczcynski, 1983; Reed & Slaichert, 1981; Reynolds, 1980) and psychiatry (Kane & Borenstein, 1985).

Sedlmeier and Gigerenzer (1989) published a paper entitled *Do studies of statistical power have an effect on the power of statistical studies?* They found that in the 25 years since Cohen's initial survey power had not changed in any substantive way. Similarly, Rossi (1990) reviewed papers published in 1982 in the *Journals of Abnormal Psychology*, *Consulting and Clinical Psychology*, and *Personality and Social Psychology*. Mean power to detect small, medium, and large effects, respectively, was 0.17, 0.57, and 0.83.

This led one of the current authors (Borenstein, 2000) to propose four theorems, as follows.

1. Power in many fields of research is abysmally low.
2. Rule (1) appears to be impervious to change.
3. The absence of significance should be interpreted as *more information is required* but is interpreted in error as meaning *no effect exists*.
4. Rule (3) appears to be impervious to change.

In a sense, then, vote counting did not originate with the narrative review. Rather, the basic mistake has existed for decades, where it found a home in primary research. When the field moved on to narrative reviews, this basic mistake was named and codified but remained basically unchanged.

There is, however, one important difference. When we are working with a single study and we have a nonsignificant result we don't have any way of knowing whether or not the effect is real. The nonsignificant p-value could reflect either the fact that the true effect is nil *or* the fact that our study had low power. While we caution against accepting the former (that the true effect is nil) we cannot rule it out.

By contrast, when we use meta-analysis to synthesize the data from a series of studies we can often identify the true effect. And in many cases (for example if the true effect is substantial and is consistent across studies) we can assert that the nonsignificant p-value in the separate studies was due to low power rather than the absence of an effect.

In the streptokinase meta-analysis on page 10, for example, it is clear that the treatment does reduce the risk of death. It is fair to say that the reason that 27 studies had nonsignificant p-values was *not* because the treatment had no effect, but rather was because of low statistical power. (In the next chapter we actually compute the power for the streptokinase studies.)

Moving beyond the null

In this chapter we have shown that *if our goal* is to test the null hypothesis, then meta-analysis (unlike the narrative review) provides a statistically sound mechanism for this purpose. However, we want to emphasize that meta-analysis allows us

to move beyond a test of the null. It allows us to assess the magnitude of the effect (which is often a more relevant question) and to determine whether or not the effect size is consistent across studies.

SUMMARY POINTS

- Vote counting is the process of counting the number of studies that are statistically significant and comparing this with the number that are not statistically significant.
- Vote counting treats a nonsignificant *p*-value as evidence that an effect is absent. In fact, though, small, moderate, and even large effect sizes may yield a nonsignificant *p*-value due to inadequate statistical power. Therefore, vote counting is never a valid approach.

unpublished, research lies dormant in the researchers' filing cabinets, and has led to the use of the term *file drawer problem* for meta-analysis.

Response

Since published studies are more likely to be included in a meta-analysis than their unpublished counterparts, there is a legitimate concern that a meta-analysis may overestimate the true effect size.

Chapter 30 (entitled *Publication Bias*) explores this question in some detail. In that chapter we discuss methods to assess the likely amount of bias in any given meta-analysis, and to distinguish between analyses that can be considered robust to the impact of publication bias from those where the results should be considered suspect.

We must remember that publication bias is a problem for any kind of literature search. The problem exists for the clinician who searches a database to locate primary studies about the utility of a treatment. It exists for persons performing a narrative review. And, it exists for persons performing a meta-analysis. Publication bias has come to be identified with meta-analysis because meta-analysis has the goal of providing a more accurate synthesis than other methods, and so we are concerned with biases that will interfere with this goal. However, it would be a mistake to conclude that this bias is not a problem for the narrative review. There, it is simply easier to ignore.

MIXING APPLES AND ORANGES

Criticism

A common criticism of meta-analysis is that researchers combine different kinds of studies (*apples and oranges*) in the same analysis. The argument is that the summary effect will ignore possibly important differences across studies.

Response

The studies that are brought together in a meta-analysis will inevitably differ in their characteristics, and the difficulty is deciding just how similar they need to be. The decision as to which studies should be included is always a judgment, and people will have different opinions on the appropriateness of combining results across studies. Some meta-analysts may make questionable judgments, and some critics may make unreasonable demands on similarity.

We need to remember that meta-analyses almost always, by their very nature, address broader questions than individual studies. Hence a meta-analysis may be thought of as asking a question about fruit, for which both apples and oranges (and indeed pears and melons) contribute valuable information. One of the strengths of meta-analysis is that the consistency, and hence generalizability, of findings from one type of study to the next can be assessed formally.

Of course, we always need to remember that we are dealing with different kinds of fruit, and to anticipate that effects may vary from one kind to the other. It is a further strength of meta-analysis that these differences, if identified, can be investigated formally. Assume, for example, that a treatment is very effective for patients with acute symptoms but has no effect for patients with chronic symptoms. If we were to combine data from studies that used both types of patients, and conclude that the treatment was modestly effective (on average), this conclusion would not be accurate for either kind of patient. If we were to restrict our attention to studies in only patients with acute symptoms, or only patients with chronic symptoms, we could report how the treatment worked with one type of patient, but could only speculate about how it would have worked with the other type. By contrast, a meta-analysis that includes data for both types of patients may allow us to address this question empirically.

GARBAGE IN, GARBAGE OUT

Criticism

The often-heard metaphor *garbage in, garbage out* refers to the notion that if a meta-analysis includes many low-quality studies, then fundamental errors in the primary studies will be carried over to the meta-analysis, where the errors may be harder to identify.

Response

Rather than thinking of meta-analysis as a process of *garbage in, garbage out* we can think of it as a process of waste management. A systematic review or meta-analysis will always have a set of inclusion criteria and these should include criteria based on the quality of the study. For trials, we may decide to limit the studies to those that use random assignment, or a placebo control. For observational studies we may decide to limit the studies to those where confounders were adequately addressed in the design or analysis. And so on. In fact, it is common in a systematic review to start with a large pool of studies and end with a much smaller set of studies after all inclusion/exclusion criteria are applied.

Nevertheless, the studies that do make it as far as a meta-analysis are unlikely to be perfect, and close attention should be paid to the possibility of bias due to study limitations. A meta-analysis of a collection of studies that is each biased in the same direction will suffer from the same bias and have higher precision. In this case, performing a meta-analysis can indeed be more dangerous than not performing one.

However, as noted in the response to the previous criticism about *apples and oranges*, a strength of meta-analysis is the ability to investigate whether variation in characteristics of studies is related to the size of the effect. Suppose that ten studies used an acceptable method to randomize patients while another ten used a questionable method. In the analysis we can compare the effect size in these two subgroups, and determine whether or not the effect size actually differs between

the two. Note that such analyses (those comparing effects in different subgroups) can have very low power so need to be interpreted carefully, especially when there are not many studies within subgroups.

IMPORTANT STUDIES ARE IGNORED

Criticism

Whereas the *garbage in, garbage out* problem relates to the inclusion of studies that perhaps should not be included, a common complementary criticism is that important studies were left out. The criticism is often leveled by people who are uncomfortable with the findings of a meta-analysis. For example, a meta-analysis to assess the effects of antioxidant supplements (beta-carotene, vitamin A, vitamin C, vitamin E, and selenium) on overall mortality was met with accusations on the web site of the Linus Pauling Institute (Oregon State University) that in this 'flawed analysis of flawed data' the authors looked at 815 human clinical trials of antioxidant supplements, but only 68 were included in the meta-analysis.

Response

We have explained that systematic reviews and meta-analyses require explicit mechanisms for deciding which studies to include and which ones to exclude. These eligibility criteria are determined by a combination of considerations of relevance and considerations of bias, and are typically decided before the search for studies is implemented. Studies should be sufficiently similar to yield results that can be interpreted, and sufficiently free of bias to yield results that can be believed. For both purposes, judgments are required, and not all meta-analysts or readers would reach the same judgments on each occasion. Importantly, in meta-analysis the criteria are transparent and are described as part of the report.

META-ANALYSIS CAN DISAGREE WITH RANDOMIZED TRIALS

Criticism

LeLorier *et al.* (1997) published a paper in which they pointed out that meta-analyses sometimes yield different results than large scale randomized trials. Specifically, they located cases in the medical literature where someone had performed a meta-analysis, and someone else subsequently performed a large scale randomized trial that addressed the same question (e.g. *Does the treatment work*?). The authors reported that the results of the meta-analysis and the randomized trial *matched* (both were statistically significant, or neither was statistically significant) in about 66% of cases, but did not match (one was statistically significant but the other was not) in the remaining 34%. Since randomized trials are generally accepted as the gold standard they conclude that some 34% of these meta-analyses were wrong, and that meta-analyses in general cannot be trusted.

Exhibit 67

WORKING TO BUILD A HEALTHY AUSTRALIA
www.nhmrc.gov.au

NHMRC levels of evidence and grades for recommendations for developers of guidelines

December 2009

NHMRC levels of evidence and grades for recommendations for developers of guidelines

Introduction

In 1999 the National Health and Medical Research Council (NHMRC) in Australia released a suite of handbooks to support organisations involved in the development of evidence-based clinical practice guidelines (www.nhmrc.gov.au/publications/synopses/cp65syn.htm).

Reflecting the general impetus of the previous decade, these handbooks focused predominantly on assessing the clinical evidence for interventions. As a consequence, the handbooks present 'levels of evidence' appropriate mainly for intervention studies. However, feedback from guideline developers received by the NHMRC indicated that the levels of evidence used by the NHMRC for intervention studies have been found to be restrictive. This was particularly so where the areas of study do not lend themselves to research designs appropriate to intervention studies (i.e. randomised controlled trials).

This paper presents a new approach to grading evidence recommendations, which should be relevant to any clinical guideline (not just those dealing with interventions).

This process of developing and grading evidence recommendations has received robust scrutiny and refinement through two public consultation phases and formal pilot-testing has been conducted across a range of guideline development projects.

The Pilot Program on 'NHMRC additional levels of evidence and grades for recommendations for developers of guidelines', was initially released for public consultation in 2005 until mid-2006 with feedback sought until 30 June 2007 on their usability and applicability. A revised version was then released for a second stage of public consultation over the period January 2008 to February 2009. Several guideline development teams, with guidance from a NHMRC Guideline Assessment Register (GAR) consultant, tested the revised grading approach in guidelines that were developed during the pilot period. The website feedback and the practical experience of guideline developers support the clinical utility and academic rigour of the new NHMRC hierarchy of levels of evidence and their role in the formulation of the new grades of recommendation.

Further peer review was solicited on one aspect of the grading process (specifically revising the levels of evidence hierarchy) through submission of a manuscript to *BMC Medical Research Methodology*, which was published in June 2009. It is anticipated that a subsequent manuscript outlining the process for grading recommendations will be submitted to a peer reviewed journal later in 2009.

Levels of evidence

Guidelines can have different purposes, dealing with clinical questions such as intervention, diagnosis, prognosis, aetiology and screening. To address these clinical questions adequately, guideline developers need to include different research designs. This consequently requires different evidence hierarchies that recognise the importance of research designs relevant to the purpose of the guideline. A new evidence hierarchy has been developed by the NHMRC GAR consultants. This hierarchy assigns levels of evidence according to the type of research question, recognising the importance of appropriate research design to that question. As well as the current NHMRC levels of evidence for interventions, new levels have been developed for studies relevant for guidelines on diagnosis, prognosis, aetiology and screening.

This consultation framework outlines the expanded levels of evidence, and provides additional information in the form of explanatory notes, a study design glossary and a summary of how the levels of evidence and other NHMRC dimensions of evidence should be used (see Part B).

Grades of recommendations

However, ascribing a level of evidence to a study, that reflects the risk of bias in its design, is only one small part of assessing evidence for a guideline recommendation. Consideration also needs to be given to: the *quality* of the study and the likelihood that the results have been affected by bias during its conduct; the *consistency* of its findings to those from other studies; the *clinical impact* of its results; the *generalisability* of the results to the population for whom the guideline is intended; and the *applicability* of the results to the Australian (and/or local) health care setting.

To further assist guideline developers to make judgments on the basis of the body of evidence relevant to a research question, a grading system for recommendations has been developed (see Part A). This takes the form of an evidence matrix, which lists the evidence components that should be considered when judging the body of evidence. The grade of a recommendation is based on an overall assessment of the rating of individual components in the evidence matrix.

Authors

This work was undertaken by the following NHMRC GAR consultants:

Revision of evidence hierarchy ("levels") -
Tracy Merlin - Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, University of Adelaide
Adele Weston - Health Technology Analysts Pty Ltd
Rebecca Tooher - ARCH: Australian Research Centre for Health of Women and Babies Division: Translational Research, Discipline of Obstetrics & Gynaecology, The University of Adelaide

Development of grading process ("grades") -
Philippa Middleton and Rebecca Tooher - ARCH: Australian Research Centre for Health of Women and Babies Division:Translational Research, Discipline of Obstetrics & Gynaecology, The University of Adelaide
Janet Salisbury – Biotext Pty Ltd.
Kristina Coleman, Sarah Norris, Adele Weston - Health Technology Analysts Pty Ltd
Karen Grimmer-Somers, Susan Hillier – Centre for Allied Health Evidence, Division of Health Sciences, University of South Australia
Tracy Merlin - Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, University of Adelaide

Acknowledgments

Feedback has been provided during this document's development phase from the following:

Marita Broadstock – New Zealand Health Technology Assessment, New Zealand
Suzanne Dyer – NHMRC Clinical Trials Centre, Australia
Paul Glasziou – Oxford University, United Kingdom
Sally Green – Australasian Cochrane Centre, Australia
Brian Haynes – McMaster University, Canada
Paul Ireland – National Institute of Clinical Studies, Australia
Nicki Jackson – Deakin University, Australia
Sally Lord and Les Irwig – University of Sydney, Australia
Skye Newton and Janet Hiller – University of Adelaide, Australia
Andrew Oxman – Oslo, Norway (GRADE Working Group)
Patricia Rogers, Professor of Public Sector Evaluation, RMIT University
Kaye Stevens, Research Fellow, RMIT University

NHMRC Management:

This project was managed by the NHMRC Evidence Translation Section with support from the NHMRC National Institute of Clinical Studies.

PART A
How to assess the body of evidence and formulate recommendations

To assist guideline developers, the NHMRC GAR consultants have developed an approach for assessing the body of evidence and formulating recommendations. This will ensure that while guidelines may differ in their purpose and formulation, their developmental processes are consistent, and their recommendations are formulated in a consistent manner. Part A describes how to grade the 'body of evidence' for each guideline recommendation. The body of evidence considers the evidence dimensions of all the studies relevant to that recommendation. Part B gives further detail on how to appraise individual studies contributing to the body of evidence.

Consequently, the NHMRC Evidence Statement Form is intended to be used for each clinical question addressed in a guideline. Before completing the form, each included study should be critically appraised and the relevant data extracted and summarised as shown in the *NHMRC standards and procedures for externally developed guidelines* (NHMRC 2007) and with reference to Part B below. This information assists in the formulation of the recommendation, and in determining the overall grade of the 'body of evidence' that supports that recommendation.

The NHMRC Evidence Statement Form sets out the basis for rating five key components of the 'body of evidence' for each recommendation. These components are:

1. The evidence base, in terms of the number of studies, level of evidence and quality of studies (risk of bias).
2. The consistency of the study results.
3. The potential clinical impact of the proposed recommendation.
4. The generalisability of the body of evidence to the target population for the guideline.
5. The applicability of the body of evidence to the Australian healthcare context.

The first two components give a picture of the internal validity of the study data in support of efficacy (for an intervention), accuracy (for a diagnostic test), or strength of association (for a prognosis or aetiological question). The third component addresses the likely clinical impact of the proposed recommendation. The last two components consider external factors that may influence the effectiveness of the proposed recommendation in practice, in terms of the generalisability of study results to the intended target population for the Guideline and setting of the proposed recommendation, and applicability to the Australian (or other local) health care system.

Definitions of the components of the evidence statement[1]

1. Evidence base

The evidence base is assessed in terms of the quantity, level and quality (risk of bias) of the included studies:

- *Quantity of evidence* reflects the number of the studies that have been included as the evidence base for each guideline (and listed in the evidence summary table or text). The quantity assessment also takes into account the number of patients in relation to the frequency of the outcomes measured (ie the statistical power of the studies). Small, underpowered studies that are otherwise sound may be included in the evidence base if their findings are generally similar — but at least some of the studies cited as evidence must be large enough to detect the size and direction of any effect. Alternatively, the results of the studies could be considered in a meta-analysis to increase the power and statistical precision of the effect estimate.

- *Level of evidence* reflects the best study types for the specific type of question (see Part B, Table 3). The most appropriate study design to answer each type of clinical question (intervention, diagnostic accuracy, aetiology or prognosis) is level II evidence. Level I studies are systematic reviews of the appropriate level II studies in each case. Study designs that are progressively less robust for answering each type of question are shown at levels III and IV. Systematic reviews of level III and IV studies are ascribed the same level of evidence as the studies included in the review to address each outcome. For example, a systematic review of cohort studies and case series for an intervention question would be given a Level III-2 ranking in the hierarchy, even if the quality of the systematic review was exceptional. The levels of evidence hierarchy is specifically concerned with the risk of bias in the presented results that is related to study design (see Explanatory note 4 to Table 3), whereas the quality of the evidence is assessed separately.

- *Quality of evidence* reflects how well the studies were conducted in order to eliminate bias, including how the subjects were selected, allocated to groups, managed and followed up and how the study outcomes were measured (see Part B, Dimensions of evidence, and Table 4 for further information).

2. Consistency

The consistency component of the 'body of evidence' assesses whether the findings are consistent across the included studies (including across a range of study populations and study designs). It is important to determine whether study results are consistent to ensure that the results are likely to be replicable or only likely to occur under certain conditions. Ideally, for a meta-analysis of randomised studies, there should be a statistical analysis of heterogeneity showing little statistical difference (consistent or homogenous) between the studies. However, given that statistical tests for heterogeneity are underpowered, presentation of an I^2 statistic[2], as well as an appraisal of the likely reasons for the differences in results across studies, would be useful. Heterogeneity in the results of studies may be due to differences in the study design, the quality of the studies (risk of bias), the population studied, the definition of the outcome being assessed, as well as many other factors. Non-randomised studies may have larger estimates of

[1] Adapted from the Scottish Intercollegiate Guidelines Network (SIGN) guide to using their Considered Judgement Form (available from http://www.sign.ac.uk/guidelines/fulltext/50/annexd.html Accessed 19.10.07)

[2] whereas most statistical tests of heterogeneity (eg Cochran's Q) assess whether heterogeneity *exists* between studies, I^2 is a statistic that quantifies *how much* heterogeneity exists between the studies (see Higgins & Thompson, 2002)

effect as a result of the greater bias in such studies; however, such studies may also be important for confirming or questioning results from randomised trials in larger populations that may be more representative of the target population for the proposed guideline.

3. Clinical impact

Clinical impact is a measure of the potential benefit from application of the guideline to a population. Factors that need to be taken into account when estimating clinical impact include:

- the relevance of the evidence to the clinical question, the statistical precision and size of the effect (including clinical importance) of the results in the evidence-base, and the relevance of the effect to the patients, compared with other management options (or none)
- the duration of therapy required to achieve the effect, and
- the balance of risks and benefits (taking into account the size of the patient population concerned).

4. Generalisability

This component covers how well the subjects and settings of the included studies will match those of the Guideline recommendations, specifically the patient population being targeted by the Guideline and the clinical setting where the recommendation will be implemented. Population issues that might influence the relative importance of recommendations include gender, age or ethnicity, baseline risk, or the level of care (eg community or hospital). This is particularly important for evidence from randomised controlled trials (RCTs), as the setting and entry requirements for such trials are generally narrowly based and therefore may not be representative of all the patients to whom the recommendation may be applied in practice. Confirmation of RCT evidence by broader-based population studies may be helpful in this regard (see '2. Consistency'). Basically, an assessment of generalisability is about determining whether the available body of evidence is answering the clinical question that was asked.

In the case of studies of diagnostic accuracy, a number of additional criteria also need to be taken into account, including the stage of the disease (eg early versus advanced), the duration of illness and the prevalence of the disease in the study population as compared to the target population for the guideline.

5. Applicability

This component addresses whether the evidence base is relevant to the Australian health care system generally, or to more local settings for specific recommendations (such as rural areas or cities).

Factors that may reduce the direct application of study findings to the Australian or more local settings include organisational factors (eg availability of trained staff, clinic time, specialised equipment, tests or other resources) and cultural factors (eg attitudes to health issues, including those that may affect compliance with the recommendation).

How to use the NHMRC Evidence Statement Form

Step 1 — Rate each of the five components

Applying evidence in real clinical situations is not usually straightforward. Consequently guideline developers find that the body of evidence supporting a recommendation rarely consists of entirely one rating for all the important components (outlined above). For example, a body of evidence may contain a large number of studies with a low risk of bias and consistent findings, but which are not directly applicable to the target population or Australian healthcare context and have only a limited clinical impact. Alternatively, a body of evidence may only consist of one or two randomised trials with small sample sizes that have a moderate risk of bias but have a very large clinical impact and are directly applicable to the Australian healthcare context and target population. The NHMRC evidence grading system is designed to allow for this mixture of components, while still reflecting the overall body of evidence supporting a guideline recommendation.

The components described above should be rated according to the matrix shown in Table 1. Enter the results into the NHMRC Evidence Statement Form (Attachment 1) along with any further notes relevant to the discussions for each component.

Table 1 Body of evidence matrix

Component	A	B	C	D
	Excellent	Good	Satisfactory	Poor
Evidence base[1]	one or more level I studies with a low risk of bias or several level II studies with a low risk of bias	one or two level II studies with a low risk of bias or a SR/several level III studies with a low risk of bias	one or two level III studies with a low risk of bias, or level I or II studies with a moderate risk of bias	level IV studies, or level I to III studies/SRs with a high risk of bias
Consistency[2]	all studies consistent	most studies consistent and inconsistency may be explained	some inconsistency reflecting genuine uncertainty around clinical question	evidence is inconsistent
Clinical impact	very large	substantial	moderate	slight or restricted
Generalisability	population/s studied in body of evidence are the same as the target population for the guideline	population/s studied in the body of evidence are similar to the target population for the guideline	population/s studied in body of evidence differ to target population for guideline but it is clinically sensible to apply this evidence to target population[3]	population/s studied in body of evidence differ to target population and hard to judge whether it is sensible to generalise to target population
Applicability	directly applicable to Australian healthcare context	applicable to Australian healthcare context with few caveats	probably applicable to Australian healthcare context with some caveats	not applicable to Australian healthcare context

SR = systematic review; several = more than two studies

[1] Level of evidence determined from the NHMRC evidence hierarchy – Table 3, Part B

[2] If there is only one study, rank this component as 'not applicable'.

[3] For example, results in adults that are clinically sensible to apply to children OR psychosocial outcomes for one cancer that may be applicable to patients with another cancer

The Evidence Statement Form also provides space to enter any other relevant factors that were taken into account by the guideline developers when judging the body of evidence and developing

the wording of the recommendation.

Step 2 — Prepare an evidence statement matrix
In the 'Evidence statement matrix ' section of the form, summarise the guideline developers' synthesis of the evidence relating to each component at the right hand side of the form, and fill in the evidence matrix at the left hand side of the form. Each recommendation should be accompanied by this matrix as well as the overall grade given to the recommendation (see Step 3). Developers should indicate dissenting opinions or other relevant issues in the space provided under the evidence matrix.

Step 3 — Formulate a recommendation based on the body of evidence
Develop wording for the recommendation. This should address the specific clinical question and ideally be written as an action statement. The wording of the recommendation should reflect the strength of the body of evidence. Words such as 'must' or 'should' are used when the evidence underpinning the recommendation is strong, and words such as 'might' or 'could' are used when the evidence base is weaker.

Step 4 — Determine the grade for the recommendation
Determine the overall grade of the recommendation based on a summation of the rating for each individual component of the body of evidence. **A recommendation cannot be graded A or B unless the evidence base and consistency of the evidence are both rated A or B.**

NHMRC overall grades of recommendation are intended to indicate the strength of the body of evidence underpinning the recommendation. This should assist users of the clinical practice guidelines to make appropriate and informed clinical judgments. Grade A or B recommendations are generally based on a body of evidence that can be trusted to guide clinical practice, whereas Grades C or D recommendations must be applied carefully to individual clinical and organisational circumstances and should be interpreted with care (see Table 2).

Table 2 Definition of NHMRC grades of recommendations

Grade of recommendation	Description
A	Body of evidence can be trusted to guide practice
B	Body of evidence can be trusted to guide practice in most situations
C	Body of evidence provides some support for recommendation(s) but care should be taken in its application
D	Body of evidence is weak and recommendation must be applied with caution

Implementing guideline recommendations
How the guideline will be implemented should be considered at the time that the guideline recommendations are being formulated. Guidelines require an implementation plan that ensures appropriate roll out, supports and evaluation of guideline effectiveness in improving practice, and guideline uptake. The Evidence Statement Form asks developers to consider four questions

related to the implementation of each recommendation:

- Will this recommendation result in changes in usual care?
- Are there any resource implications associated with implementing this recommendation?
- Will the implementation of this recommendation require changes in the way care is currently organised?
- Are the guideline development group aware of any barriers to the implementation of this recommendation?

NHMRC Evidence Statement

(If rating is not completely clear, use the space next to each criteria to note how the group came to a judgment. Part B of this document will assist with the critical appraisal of individual studies included in the body of evidence)

Key question(s):		**Evidence table ref:**
1. Evidence base *(number of studies, level of evidence and risk of bias in the included studies)*		
	A	**One or more level I studies with a low risk of bias or several level II studies with a low risk of bias**
	B	**One or two Level II studies with a low risk of bias or SR/several Level III studies with a low risk of bias**
	C	**One or two Level III studies with a low risk of bias or Level I or II studies with a moderate risk of bias**
	D	**Level IV studies or Level I to III studies/SRs with a high risk of bias**
2. Consistency *(if only one study was available, rank this component as 'not applicable')*		
	A	**All studies consistent**
	B	**Most studies consistent and inconsistency can be explained**
	C	**Some inconsistency, reflecting genuine uncertainty around question**
	D	**Evidence is inconsistent**
	NA	**Not applicable (one study only)**
3. Clinical impact *(Indicate in the space below if the study results varied according to some unknown factor (not simply study quality or sample size) and thus the clinical impact of the intervention could not be determined)*		
	A	**Very large**
	B	**Moderate**
	C	**Slight**
	D	**Restricted**
4. Generalisability *(How well does the body of evidence match the population and clinical settings being targeted by the Guideline?)*		
	A	**Evidence directly generalisable to target population**
	B	**Evidence directly generalisable to target population with some caveats**
	C	**Evidence not directly generalisable to the target population but could be sensibly applied**
	D	**Evidence not directly generalisable to target population and hard to judge whether it is sensible to apply**
5. Applicability *(Is the body of evidence relevant to the Australian healthcare context in terms of health services/delivery of care and cultural factors?)*		
	A	**Evidence directly applicable to Australian healthcare context**
	B	**Evidence applicable to Australian healthcare context with few caveats**
	C	**Evidence probably applicable to Australian healthcare context with some caveats**
	D	**Evidence not applicable to Australian healthcare context**

Other factors *(Indicate here any other factors that you took into account when assessing the evidence base (for example, issues that might cause the group to downgrade or upgrade the recommendation)*

EVIDENCE STATEMENT MATRIX

Please summarise the development group's synthesis of the evidence relating to the key question, taking all the above factors into account.

Component	Rating	Description
1. Evidence base		
2. Consistency		
3. Clinical impact		
4. Generalisability		
5. Applicability		

Indicate any dissenting opinions

RECOMMENDATION	GRADE OF RECOMMENDATION	
What recommendation(s) does the guideline development group draw from this evidence? Use action statements where possible.		

UNRESOLVED ISSUES *If needed, keep note of specific issues that arise when each recommendation is formulated and that require follow-up.*

IMPLEMENTATION OF RECOMMENDATION *Please indicate yes or no to the following questions. Where the answer is yes please provide explanatory information about this. This information will be used to develop the implementation plan for the guidelines.*	
Will this recommendation result in changes in usual care?	**YES**
	NO
Are there any resource implications associated with implementing this recommendation?	**YES**
	NO
Will the implementation of this recommendation require changes in the way care is currently organised?	**YES**
	NO
Are the guideline development group aware of any barriers to the implementation of this recommendation?	**YES**
	NO

PART B

Implementing NHMRC dimensions of evidence including the new levels of evidence hierarchy

This part of the document outlines how individual studies included in a systematic literature review should be assessed using the NHMRC dimensions of evidence and provides levels of evidence appropriate for the most common types of research questions. The basic principles of systematic reviewing and assessing evidence are set out in the NHMRC handbook series on the development of clinical practice guidelines (NHMRC 2000ab).

Dimensions of evidence for assessing included studies

Each included study in a systematic review should be assessed according to the following three dimensions of evidence:

1. Strength of evidence

a. *Level of evidence*: Each study design is assessed according to its place in the research hierarchy. The hierarchy reflects the potential of each study or systematic review included in the systematic review(s) underpinning the Guidelines to adequately answer a particular research question, based on the probability that its design has minimised the impact of bias on the results. See page 6–10 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).
The original NHMRC levels of evidence for intervention studies (NHMRC 2000b), together with the new levels of evidence for questions on diagnosis, prognosis, aetiology and screening are shown in the evidence hierarchy in Table 3. A glossary describing each of the study designs is provided in Attachment 2.

b. *Quality of evidence* (risk of bias): The methodological quality of each included study is critically appraised. Each study is assessed according to the likelihood that bias, confounding and/or chance may have influenced its results. The NHMRC toolkit *How to review the evidence: systematic identification and review of the scientific literature* (NHMRC 2000a) lists examples of ways that methodological quality can be assessed. In cases where other critical appraisal approaches may be required, there are a number of alternatives. The NHMRC/NICS can advise on the choice of an alternative to supplement and/or replace those in the NHMRC handbook (see Table 4).

c. *Statistical precision*: The primary outcomes of each included study are evaluated to determine whether the effect is real, rather than due to chance (using a level of significance expressed as a *P*-value and/or a confidence interval). See page 17 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).

2. Size of effect

This dimension is useful for assessing the clinical importance of the findings of each study (and hence addresses the clinical impact component of the body of evidence matrix in Part A). This is a different concept to statistical precision and specifically refers to the measure of effect or point estimate provided in the results of each study (eg mean difference, relative risk, odds ratio, hazard ratio, sensitivity, specificity). In the case of a meta-analysis it is the pooled measure of effect from the studies included in the systematic review (eg weighted mean difference, pooled relative risk). These point estimates are calculated in comparison to either doing nothing or versus an active control.

Size of the effect therefore refers to the distance of the point estimate from its null value for each outcome (or result) and the values included in the corresponding 95% confidence interval. For example, for a ratio such as a relative risk the null value is 1.0 and so a relative of risk of 5 is a large point estimate; for a mean difference the null value is zero (indicating no difference) and so a mean difference of 1.5kg may be small. The size of the effect indicates just how much clinical impact that particular factor or intervention will have on the patient and should always be taken in the context of what is a clinically relevant difference for the patient. The upper and lower point estimates in the confidence interval can then be used to judge whether it is likely that most of the time the intervention will have a clinically important impact, or that it is possible that in some instances the impact will be clinically unimportant or that there will be no impact. See pages 17–23 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).

3. Relevance of evidence

This dimension deals with the translation of research evidence into clinical practice and is potentially the most subjective of the evidence assessments. There are two key questions.

a. *Appropriateness of the outcomes*: Are the outcomes measured in the study relevant to patients? This question focuses on the patient-centredness of the study. See pages 23–27 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).
b. *Relevance of study question*: How closely do the elements of the research question ('PICO'[3]) match those of the clinical question being considered in the guideline? This is important in determining the extent to which the study results are relevant (generalisable) for the population who will be the recipients of the clinical guideline.

The results of these assessments for each included study should be entered into a data extraction form described in the *NHMRC standards and procedures for externally developed guidelines* (NHMRC 2007). Once each included study is assessed according to these dimensions of evidence, a summary can be made that is relevant to the whole body of evidence, which can then be graded as described in Part A of this document. The data extraction process provides the evidence base on which the systematic review, and subsequent guideline recommendations are built.

[3] P=Population, I=Intervention/index test/indicator, C=Comparison, O=Outcome

Table 3 NHMRC Evidence Hierarchy: designations of 'levels of evidence' according to type of research question (including explanatory notes)

Level	Intervention [1]	Diagnostic accuracy [2]	Prognosis	Aetiology [3]	Screening Intervention
I [4]	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies
II	A randomised controlled trial	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among consecutive persons with a defined clinical presentation[6]	A prospective cohort study[7]	A prospective cohort study	A randomised controlled trial
III-1	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among non-consecutive persons with a defined clinical presentation[6]	All or none[8]	All or none[8]	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)
III-2	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial[9] ▪ Cohort study ▪ Case-control study ▪ Interrupted time series with a control group	A comparison with reference standard that does not meet the criteria required for Level II and III-1 evidence	Analysis of prognostic factors amongst persons in a single arm of a randomised controlled trial	A retrospective cohort study	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial ▪ Cohort study ▪ Case-control study
III-3	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study[10] ▪ Interrupted time series without a parallel control group	Diagnostic case-control study[6]	A retrospective cohort study	A case-control study	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study
IV	Case series with either post-test or pre-test/post-test outcomes	Study of diagnostic yield (no reference standard)[11]	Case series, or cohort study of persons at different stages of disease	A cross-sectional study or case series	Case series

Explanatory notes

[1] Definitions of these study designs are provided on pages 7-8 *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b*)* and in the accompanying Glossary.

[2] These levels of evidence apply only to studies of assessing the accuracy of diagnostic or screening tests. To assess the overall effectiveness of a diagnostic test there also needs to be a consideration of the impact of the test on patient management and health outcomes (Medical Services Advisory Committee 2005, Sackett and Haynes 2002). The evidence hierarchy given in the 'Intervention' column should be used to assess the impact of a diagnostic test on health outcomes relative to an existing method of diagnosis/comparator test(s). The evidence hierarchy given in the 'Screening' column should be used to assess the impact of a screening test on health outcomes relative to no screening or opportunistic screening.

[3] If it is possible and/or ethical to determine a causal relationship using experimental evidence, then the 'Intervention' hierarchy of evidence should be utilised. If it is only possible and/or ethical to determine a causal relationship using observational evidence (eg. cannot allocate groups to a potential harmful exposure, such as nuclear radiation), then the 'Aetiology' hierarchy of evidence should be utilised.

[4] A systematic review will only be assigned a level of evidence as high as the studies it contains, excepting where those studies are of level II evidence. Systematic reviews of level II evidence provide more data than the individual studies and any meta-analyses will increase the precision of the overall results, reducing the likelihood that the results are affected by chance. Systematic reviews of lower level evidence present results of likely poor internal validity and thus are rated on the likelihood that the results have been affected by bias, rather than whether the systematic review itself is of good quality. Systematic review *quality* should be assessed separately. A systematic review should consist of at least two studies. In systematic reviews that include different study designs, the overall level of evidence should relate to each individual outcome/result, as different studies (and study designs) might contribute to each different outcome.

[5] The validity of the reference standard should be determined in the context of the disease under review. Criteria for determining the validity of the reference standard should be pre-specified. This can include the choice of the reference standard(s) and its timing in relation to the index test. The validity of the reference standard can be determined through quality appraisal of the study (Whiting et al 2003).

[6] Well-designed population based case-control studies (eg. population based screening studies where test accuracy is assessed on all cases, with a random sample of controls) do capture a population with a representative spectrum of disease and thus fulfil the requirements for a valid assembly of patients. However, in some cases the population assembled is not representative of the use of the test in practice. In diagnostic case-control studies a selected sample of patients already known to have the disease are compared with a separate group of normal/healthy people known to be free of the disease. In this situation patients with borderline or mild expressions of the disease, and conditions mimicking the disease are excluded, which can lead to exaggeration of both sensitivity and specificity. This is called spectrum bias or spectrum effect because the spectrum of study participants will not be representative of patients seen in practice (Mulherin and Miller 2002).

[7] At study inception the cohort is either non-diseased or all at the same stage of the disease. A randomised controlled trial with persons either non-diseased or at the same stage of the disease in *both* arms of the trial would also meet the criterion for this level of evidence.

[8] All or none of the people with the risk factor(s) experience the outcome; and the data arises from an unselected or representative case series which provides an unbiased representation of the prognostic effect. For example, no smallpox develops in the absence of the specific virus; and clear proof of the causal link has come from the disappearance of small pox after large-scale vaccination.

[9] This also includes controlled before-and-after (pre-test/post-test) studies, as well as adjusted indirect comparisons (ie. utilise A vs B and B vs C, to determine A vs C with statistical adjustment for B).

[10] Comparing single arm studies ie. case series from two studies. This would also include unadjusted indirect comparisons (ie. utilise A vs B and B vs C, to determine A vs C but where there is no statistical adjustment for B).

[11] Studies of diagnostic yield provide the yield of diagnosed patients, as determined by an index test, without confirmation of the accuracy of this diagnosis by a reference standard. These may be the only alternative when there is no reliable reference standard.

Note A: Assessment of comparative harms/safety should occur according to the hierarchy presented for each of the research questions, with the proviso that this assessment occurs within the context of the topic being assessed. Some harms (and other outcomes) are rare and cannot feasibly be captured within randomised controlled trials, in which case lower levels of evidence may be the only type of evidence that is practically achievable; physical harms and psychological harms may need to be addressed by different study designs; harms from diagnostic testing include the likelihood of false positive and false negative results; harms from screening include the likelihood of false alarm and false reassurance results.

Note B: When a level of evidence is attributed in the text of a document, it should also be framed according to its corresponding research question eg. level II intervention evidence; level IV diagnostic evidence; level III-2 prognostic evidence.

Note C: Each individual study that is attributed a "level of evidence" should be rigorously appraised using validated or commonly used checklists or appraisal tools to ensure that factors other than study design have not affected the validity of the results.

Source: Hierarchies adapted and modified from: NHMRC 1999; Bandolier 1999; Lijmer et al. 1999; Phillips et al. 2001.

Table 4 Assessment of individual study quality

Study type	Location of NHMRC checklist[1]	Additional/supplemental quality assessment tool
Intervention	Page 45	
Diagnosis	Page 62	QUADAS (Whiting et al., 2003)
Prognosis	Page 81	GATE checklist for prognostic studies (NZGG, 2001)
Aetiology	Page 73	
Screening	Page 45	UK National Screening Committee Guidelines (2000)
Systematic Review	Page 16[2]	SIGN checklist (SIGN, 2006), CASP checklist (CASP, 2006)

[1] Included in *How to review the evidence: systematic identification and review of the scientific literature* (NHMRC 2000a)

[2] Included in *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b)

Conclusion

This paper outlines an approach to developing guideline recommendations that was piloted and refined over four years by NHMRC GAR consultants. This approach reflects the concerted input of experience in assisting a range of guideline developers to develop guidelines for a range of conditions and purposes. It also incorporates feedback from the guideline developers themselves to improve the utility of the process and the clarity of the instructions and suggestions.

There are some types of evidence that have not been captured in this new grading approach, specifically the appraisal of qualitative studies and cost-effectiveness analyses. The empirical and theoretical basis for appraising and synthesising these types of evidence in a standard manner is still uncertain and undergoing refinement. It is expected that that with developments in these fields that subsequent revision of the presented approach to developing guideline recommendations may occur.

This new methodological approach provides a way forward for guideline developers to appraise, classify and grade evidence relevant to the purpose of a guideline and develop recommendations that are evidence-based, action-oriented and implementable.

ATTACHMENT 2

STUDY DESIGN GLOSSARY (alphabetic order)

Adapted from NHMRC 2000ab, Glasziou et al 2001, Elwood 1998

Note: This is a specialised glossary that relates specifically to the study designs mentioned in the NHMRC Evidence Hierarchy. Glossaries of terms that relate to wider epidemiological concepts and evidence based medicine are also available – see http://www.inahta.org/HTA/Glossary/; http://www.ebmny.org/glossary.html

All or none –- all or none of a series of people (case series) with the risk factor(s) experience the outcome. The data should relate to an unselected or representative case series which provides an unbiased representation of the prognostic effect. For example, no smallpox develops in the absence of the specific virus; and clear proof of the causal link has come from the disappearance of small pox after large scale vaccination. This is a rare situation.

A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among consecutive patients with a defined clinical presentation – a cross-sectional study where a consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among non-consecutive patients with a defined clinical presentation – a cross-sectional study where a non-consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

Adjusted indirect comparisons – an adjusted indirect comparison compares single arms from two or more interventions from two or more separate studies via the use of a common reference ie A versus B and B versus C allows a comparison of A versus C when there is statistical adjustment for B. This is most commonly done in meta-analyses (see Bucher et al 1997). Such an indirect comparison should only be attempted when the study populations, common comparator/reference, and settings are very similar in the two studies (Song et al 2000).

Case-control study – people with the outcome or disease (cases) and an appropriate group of controls without the outcome or disease (controls) are selected and information obtained about their previous exposure/non-exposure to the intervention or factor under study.

Case series – a single group of people exposed to the intervention (factor under study).

> **Post-test** – only outcomes after the intervention (factor under study) are recorded in the series of people, so no comparisons can be made.
>
> **Pre-test/post-test** – measures on an outcome are taken before and after the intervention is introduced to a series of people and are then compared (also known as a 'before- and-after study').

Cohort study – outcomes for groups of people observed to be exposed to an intervention, or the factor under study, are compared to outcomes for groups of people not exposed.

Prospective cohort study – where groups of people (cohorts) are observed at a point in time to be exposed or not exposed to an intervention (or the factor under study) and then are followed prospectively with further outcomes recorded as they happen.

Retrospective cohort study – where the cohorts (groups of people exposed and not exposed) are defined at a point of time in the past and information collected on subsequent outcomes, eg. the use of medical records to identify a group of women using oral contraceptives five years ago, and a group of women not using oral contraceptives, and then contacting these women or identifying in subsequent medical records the development of deep vein thrombosis.

Cross-sectional study – a group of people are assessed at a particular point (or cross-section) in time and the data collected on outcomes relate to that point in time ie proportion of people with asthma in October 2004. This type of study is useful for hypothesis-generation, to identify whether a risk factor is associated with a certain type of outcome, but more often than not (except when the exposure and outcome are stable eg. genetic mutation and certain clinical symptoms) the causal link cannot be proven unless a time dimension is included.

Diagnostic (test) accuracy – in diagnostic accuracy studies, the outcomes from one or more diagnostic tests under evaluation (the *index test*/s) are compared with outcomes from a *reference standard test*. These outcomes are measured in individuals who are suspected of having the condition of interest. The term *accuracy* refers to the amount of agreement between the index test and the reference standard test in terms of outcome measurement. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver operator characteristic (ROC) curve (Bossuyt et al 2003)

Diagnostic case-control study – the index test results for a group of patients already known to have the disease (through the reference standard) are compared to the index test results with a separate group of normal/healthy people known to be free of the disease (through the use of the reference standard). In this situation patients with borderline or mild expressions of the disease, and conditions mimicking the disease are excluded, which can lead to exaggeration of both sensitivity and specificity. This is called spectrum bias because the spectrum of study participants will not be representative of patients seen in practice. *Note: this does not apply to well-designed population based case-control studies.*

Historical control study – outcomes for a prospectively collected group of people exposed to the intervention (factor under study) are compared with either (1) the outcomes of people treated at the same institution prior to the introduction of the intervention (ie. control group/usual care), or (2) the outcomes of a previously published series of people undergoing the alternate or control intervention.

Interrupted time series with a control group – trends in an outcome or disease are measured over multiple time points before and after the intervention (factor under study) is introduced to a group of people, and then compared to the outcomes at the same time points for a group of people that do not receive the intervention (factor under study).

Interrupted time series without a parallel control group – trends in an outcome or disease are measured over multiple time points before and after the intervention (factor under study) is introduced to a group of people, and compared (as opposed to being compared to an external control group).

Non-randomised, experimental trial - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention group or a control group, using a non-random method (such as patient or clinician preference/availability) and the outcomes from each group are compared.

This can include:

(1) **a controlled before-and-after study**, where outcome measurements are taken before and after the intervention is introduced, and compared at the same time point to outcome measures in the (control) group.

(2) **an adjusted indirect comparison**, where two randomised controlled trials compare different interventions to the same comparator ie. the placebo or control condition. The outcomes from the two interventions are then compared indirectly. *See entry on adjusted indirect comparisons.*

Pseudo-randomised controlled trial - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention (the factor under study) group or a control group, using a pseudo-random method (such as alternate allocation, allocation by days of the week or odd-even study numbers) and the outcomes from each group are compared.

Randomised controlled trial – the unit of experimentation (eg. people, or a cluster of people[4]) is allocated to either an intervention (the factor under study) group or a control group, using a random mechanism (such as a coin toss, random number table, computer-generated random numbers) and the outcomes from each group are compared. Cross-over randomised controlled trials – where the people in the trial receive one intervention and then cross-over to receive the alternate intervention at a point in time – are considered to be the same level of evidence as a randomised controlled trial, although appraisal of these trials would need to be tailored to address the risk of bias specific to cross-over trials,

Reference standard - the reference standard is considered to be the best available method for establishing the presence or absence of the target condition of interest. The reference standard can be a single method, or a combination of methods. It can include laboratory tests, imaging tests, and pathology, but also dedicated clinical follow-up of individuals (Bossuyt et al 2003).

Screening intervention – a screening intervention is a public health service in which members of a defined population, who do not necessarily perceive that they are at risk of, or are already affected by a disease or its complications (asymptomatic), are asked a question or offered a test, to identify those individuals who are more likely to be helped than harmed by further tests or treatment to reduce the risk of a disease or its complications (UK National Screening Committee, 2007). A screening intervention study compares the implementation of the screening intervention in an asymptomatic population with a control group where the screening intervention is not employed or where a different screening intervention is employed. The aim is to see whether the screening intervention of interest results in improvements in patient-relevant outcomes eg survival.

Study of diagnostic yield – these studies provide the yield of diagnosed patients, as determined by the index test, without confirmation of the accuracy of the diagnosis (ie. whether the patient is actually diseased) by a reference standard test.

Systematic review – systematic location, appraisal and synthesis of evidence from scientific studies.

[4] Known as a cluster randomised controlled trial

Test - any method of obtaining additional information on a person's health status. It includes information from history and physical examination, laboratory tests, imaging tests, function tests, and histopathology (Bossuyt et al 2003).

Two or more single arm study – the outcomes of a single series of people receiving an intervention (case series) from two or more studies are compared. *Also see entry on unadjusted indirect comparisons.*

Unadjusted indirect comparisons – an unadjusted indirect comparison compares single arms from two or more interventions from two or more separate studies via the use of a common reference ie A versus B and B versus C allows a comparison of A versus C but there is no statistical adjustment for B. Such a simple indirect comparison is unlikely to be reliable (see Song et al 2000).

References

Bandolier editorial. Diagnostic testing emerging from the gloom? *Bandolier*, 1999;70. Available at: http://www.jr2.ox.ac.uk/bandolier/band70/b70-5.html

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HCW for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *AJR*, 2003; 181:51-56

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*, 1997;50:683-91.

CASP (2006). Critical Appraisal Skills Programme (CASP) - making sense of evidence: 10 questions to help you make sense of reviews. England: Public Health Resource Unit. Available at: http://www.phru.nhs.uk/Doc_Links/S.Reviews%20Appraisal%20Tool.pdf

Elwood M. (1998) *Critical appraisal of epidemiological studies and clinical trials*. Second edition. Oxford: Oxford University Press.

Glasziou P, Irwig L, Bain C, Colditz G. (2001) *Systematic reviews in health care. A practical guide.* Cambridge: Cambridge University Press.

Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*, 2002; 21(11):1539-58.

Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 1999; 282(11):1061-6.

Medical Services Advisory Committee (2005). *Guidelines for the assessment of diagnostic technologies*. [Internet] Available at: www.msac.gov.au

Mulherin S, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med*, 2002;137:598-602.

NHMRC (1999). *A guide to the development, implementation and evaluation of clinical practice guidelines*. Canberra: National Health and Medical Research Council.

NHMRC (2000a). *How to review the evidence: systematic identification and review of the scientific literature*. Canberra: National Health and Medical Research Council.

NHMRC (2000b). *How to use the evidence: assessment and application of scientific evidence*. Canberra: National Health and Medical Research Council.

NHMRC (2007). *NHMRC standards and procedures for externally developed guidelines*. Canberra: National Health and Medical Research Council.
http://www.nhmrc.gov.au/publications/synopses/_files/nh56.pdf

NZGG (2001). *Handbook for the preparation of explicit evidence-based clinical practice guidelines*. Wellington: New Zealand Guidelines Group. Available at: http://www.nzgg.org.nz

Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M (2001). *Oxford Centre for Evidence-Based Medicine levels of evidence (May 2001).* Oxford: Centre for Evidence-Based Medicine. Available at: http://www.cebm.net/levels_of_evidence.asp

Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ*, 2002;324:539-41.

SIGN. *SIGN 50. A guideline developers' handbook*. Methodology checklist 1: Systematic reviews and meta-analyses. Edinburgh: Scottish Intercollegiate Guidelines Network. Available at: http://www.sign.ac.uk/guidelines/fulltext/50/checklist1.html

Song F, Glenny A-M, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. *Controlled Clinical Trials*, 2000;21(5):488-497.

UK National Screening Committee (2000). *The UK National Screening Committee's criteria for appraising the viability, effectiveness and appropriateness of a screening programme.* In: Second Report of the UK National Screening Committee. London: United Kingdom Departments of Health. Pp. 26-27. Available at: http://www.nsc.nhs.uk/

UK National Screening Committee. *What is screening?*. [Internet]. Available at - http://www.nsc.nhs.uk/whatscreening/whatscreen_ind.htm [Accessed August 2007].

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS:a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; 3(1): 25. Available at: http://www.biomedcentral.com/1471-2288/3/25

Exhibit 68

Published in final edited form as:
Plast Reconstr Surg. 2011 July ; 128(1): 305–310. doi:10.1097/PRS.0b013e318219c171.

The Levels of Evidence and their role in Evidence-Based Medicine

Patricia B. Burns, MPH[1], **Rod J. Rohrich, MD**[2], and **Kevin C. Chung, MD, MS**[3]

[1]Research Associate, Section of Plastic Surgery, Department of Surgery, The University of Michigan Health System

[2]Professor of Surgery, Department of Plastic Surgery, University of Texas Southwestern Medical Center

[3]Professor of Surgery, Section of Plastic Surgery, Department of Surgery, The University of Michigan Health System

Abstract

As the name suggests, evidence-based medicine (EBM), is about finding evidence and using that evidence to make clinical decisions. A cornerstone of EBM is the hierarchical system of classifying evidence. This hierarchy is known as the levels of evidence. Physicians are encouraged to find the highest level of evidence to answer clinical questions. Several papers published in Plastic Surgery journals concerning EBM topics have touched on this subject.[1–6] Specifically, previous papers have discussed the lack of higher level evidence in PRS and need to improve the evidence published in the journal. Before that can be accomplished, it is important to understand the history behind the levels and how they should be interpreted. This paper will focus on the origin of levels of evidence, their relevance to the EBM movement and the implications for the field of plastic surgery as well as the everyday practice of plastic surgery.

Keywords

Evidence-based medicine; levels of evidence

History of Levels of Evidence

The levels of evidence were originally described in a report by the Canadian Task Force on the Periodic Health Examination in 1979.[7] The report's purpose was to develop recommendations on the periodic health exam and base those recommendations on evidence in the medical literature. The authors developed a system of rating evidence (Table 1) when determining the effectiveness of a particular intervention. The evidence was taken into account when grading recommendations. For example, a Grade A recommendation was given if there was good evidence to support a recommendation that a condition be included in the periodic health exam. The levels of evidence were further described and expanded by Sackett[8] in an article on levels of evidence for antithrombotic agents in 1989 (Table 2). Both

Corresponding Author: Kevin C. Chung, MD, MS, Section of Plastic Surgery, University of Michigan Health System, 2130 Taubman Center, SPC 5340, 1500 E. Medical Center Drive, Ann Arbor, MI, 48109-5340, kecchung@umich.edu, Phone 734-936-5885, Fax 734-763-5354.

systems place randomized controlled trials (RCT) at the highest level and case series or expert opinions at the lowest level. The hierarchies rank studies according to the probability of bias. RCTs are given the highest level because they are designed to be unbiased and have less risk of systematic errors. For example, by randomly allocating subjects to two or more treatment groups, these types of studies also randomize confounding factors that may bias results. A case series or expert opinion is often biased by the author's experience or opinions and there is no control of confounding factors.

Modification of levels

Since the introduction of levels of evidence, several other organizations and journals have adopted variation of the classification system. Diverse specialties are often asking different questions and it was recognized that the type and level of evidence needed to be modified accordingly. Research questions are divided into the categories: treatment, prognosis, diagnosis, and economic/decision analysis. For example, Table 3 shows the levels of evidence developed by the American Society of Plastic Surgeons (ASPS) for prognosis[9] and Table 4 shows the levels developed by the Centre for Evidence Based Medicine (CEBM) for treatment.[10] The two tables highlight the types of studies that are appropriate for the question (prognosis versus treatment) and how quality of data is taken into account when assigning a level. For example, RCTs are not appropriate when looking at the prognosis of a disease. The question in this instance is: "What will happen if we do nothing at all"? Because a prognosis question does not involve comparing treatments, the highest evidence would come from a cohort study or a systematic review of cohort studies. The levels of evidence also take into account the quality of the data. For example, in the chart from CEBM, poorly designed RCTs have the same level of evidence as a cohort study.

A grading system that provides strength of recommendations based on evidence has also changed over time. Table 5 shows the Grade Practice Recommendations developed by ASPS. The grading system provides an important component in evidence-based medicine and assists in clinical decision making. For example, a strong recommendation is given when there is level I evidence and consistent evidence from Level II, III and IV studies available. The grading system does not degrade lower level evidence when deciding recommendations if the results are consistent.

Interpretation of levels

Many journals assign a level to the papers they publish and authors often assign a level when submitting an abstract to conference proceedings. This allows the reader to know the level of evidence of the research but the designated level of evidence does always guarantee the quality of the research. It is important that readers not assume that level 1 evidence is always the best choice or appropriate for the research question. This concept will be very important for all of us to understand as we evolve into the field of EBM in Plastic Surgery. By design, our designated surgical specialty will always have important articles that may have a lower level of evidence due to the level of innovation and technique articles which are needed to move our surgical specialty forward.

Although RCTs are the often assigned the highest level of evidence, not all RCTs are conducted properly and the results should be carefully scrutinized. Sackett [8] stressed the importance of estimating types of errors and the power of studies when interpreting results from RCTs. For example, a poorly conducted RCT may report a negative result due to low power when in fact a real difference exists between treatment groups. Scales such as the Jadad scale have been developed to judge the quality of RCTs.[11] Although physicians may not have the time or inclination to use a scale to assess quality, there are some basic items that should be taken into account. Items used for assessing RCTs include: randomization,

blinding, a description of the randomization and blinding process, description of the number of subjects who withdrew or drop out of the study; the confidence intervals around study estimates; and a description of the power analysis. For example, Bhandari et al.[12] published a paper assessing the quality of surgical RCTs. The authors evaluated the quality of RCTs reported in the Journal of Bone and Joint Surgery (JBJS) from 1988–2000. Papers with a score of > 75% were deemed high quality and 60% of the papers had a score < 75%. The authors identified 72 RCTs during this time period and the mean score was 68%. The main reason for the low quality score was lack of appropriate randomization, blinding, and a description of patient exclusion criteria. Another paper found the same quality score of papers in JBJS with a level 1 rating compared to level 2. [13] Therefore, one should not assume that level 1 studies have higher quality than level 2.

A resource for surgeons when appraising levels of evidence are the users' guides published in the Canadian Journal of Surgery[14, 15] and the Journal of Bone and Joint Surgery.[16] Similar papers that are not specific to surgery have been published in the Journal of the American Medical Association (JAMA).[17, 18]

Plastic surgery and EBM

The field of plastic surgery has been slow to adopt evidence-based medicine. This was demonstrated in a paper examining the level of evidence of papers published in PRS.[19] The authors assigned levels of evidence to papers published in PRS over a 20 year period. The majority of studies (93% in 1983) were level 4 or 5, which denotes case series and case reports. Although the results are disappointing, there was some improvement over time. By 2003 there were more level 1studies (1.5%) and fewer level 4 and 5 studies (87%). A recent analysis looked at the number of level 1 studies in 5 different plastic surgery journals from 1978–2009. The authors defined level 1 studies as RCTs and meta-analysis and restricted their search to these studies. The number of level 1 studies increased from 1 in 1978 to 32 by 2009.[20] From these results, we see that the field of plastic surgery is improving the level of evidence but still has a way to go, especially in improving the quality of studies published. For example, approximately a third of the studies involved double blinding, but the majority did not randomize subjects, describe the randomization process, or perform a power analysis. Power analysis is another area of concern in plastic surgery. A review of the plastic surgery literature found that the majority of published studies have inadequate power to detect moderate to large differences between treatment groups.[21] No matter what the level of evidence for a study, if it is under powered, the interpretation of results is questionable.

Although the goal is to improve the overall level of evidence in plastic surgery, this does not mean that all lower level evidence should be discarded. Case series and case reports are important for hypothesis generation and can lead to more controlled studies. Additionally, in the face of overwhelming evidence to support a treatment, such as the use of antibiotics for wound infections, there is no need for an RCT.

Clinical examples using levels of evidence

In order to understand how the levels of evidence work and aid the reader in interpreting levels, we provide some examples from the plastic surgery literature. The examples also show the peril of medical decisions based on results from case reports.

An association was hypothesized between lymphoma and silicone breast implants based on case reports.[22–27] The level of evidence for case reports, depending on the scale used, is 4 or 5. These case reports were used to generate the hypothesis that a possible association existed. Because of these results, several large retrospective cohort studies from the United States, Canada, Denmark, Sweden and Finland were conducted.[28–32] The level of evidence

for a retrospective cohort is 2. All of these studies had many years of follow-up for a large number of patients. Some of the studies found an elevated risk and others no risk for lymphoma. None of the studies reached statistical significance. Therefore, higher level evidence from cohort studies does not provide evidence of any risk of lymphoma. Finally, a systematic review was performed that combined the evidence from the retrospective cohorts.[27] The results found an overall standardized incidence ratio of 0.89 (95% CI 0.67–1.18). Because the confidence intervals include 1, the results indicate there is no increased incidence. The level of evidence for the systematic review is 1. Based on the best available evidence, there is no association between lymphoma and silicone implants. This example shows how low level evidence studies were used to generate a hypothesis, which then led to higher level evidence that disproved the hypothesis. This example also demonstrates that RCTs are not feasible for rare events such as cancer and the importance of observational studies for a specific study question. A case-control study is a better option and provides higher evidence for testing the prognosis of the long-term effect of silicone breast implants.

Another example is the injection of epinephrine in fingers. Based on case reports prior to 1950, physicians were advised that epinephrine injection can result in finger ischemia.[33] We see in this example in which level 4 or 5 evidence was accepted as fact and incorporated into medical textbooks and teaching. However, not all physicians accepted this evidence and are performing injections of epinephrine into the fingers with no adverse effects on the hand. Obviously, it was time for higher level evidence to resolve this issue. An in-depth review of the literature from 1880 to 2000 by Denkler,[33] identified 48 cases of digital infarction of which 21 were injected with epinephrine. Further analysis found that the addition of procaine to the epinephrine injection was the cause of the ischemia.[34] The procaine used in these injections included toxic acidic batches that were recalled in 1948. In addition, several cohort studies found no complications from the use of epinephrine in the fingers and hand.[3536, 37] The results from these cohort studies increased the level of evidence. Based on the best available evidence from these studies, the hypothesis that epinephrine injection will harm fingers was rejected. This example highlights the biases inherent in case reports. It also shows the risk when spurious evidence is handed down and integrated into medical teaching.

Obtaining the best evidence

We have established the need for RCTs to improve evidence in plastic surgery but have also acknowledged the difficulties, particularly with randomization and blinding. Although RCTs may not be appropriate for many surgical questions, well designed and conducted cohort or case-control studies could boost the level of evidence. Many of the current studies tend to be descriptive and lack a control group. The way forward seems clear. Plastic surgery researchers need to consider utilizing a cohort or case-control design whenever an RCT is not possible. If designed properly, the level of evidence for observational studies can approach or surpass those from an RCT. In some instances, observation studies and RCTs have found similar results.[38] If enough cohort or case-control studies become available, this increases the prospect of systematic reviews of these studies that will increase overall evidence levels in plastic surgery.

Conclusion

The levels of evidence are an important component of EBM. Understanding the levels and why they are assigned to publications and abstracts helps the reader to prioritize information. This is not to say that all level 4 evidence should be ignored and all level 1 evidence accepted as fact. The levels of evidence provide a guide and the reader needs to be cautious when interpreting these results.

Acknowledgments

Supported in part by a Midcareer Investigator Award in Patient-Oriented Research (K24 AR053120) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (to Dr. Kevin C. Chung).

References

1. McCarthy CM, Collins ED, Pusic AL. Where do we find the best evidence? Plast Reconstr Surg. 2008; 122:1942–1947. [PubMed: 19050548]
2. Chung KC, Swanson JA, Schmitz D, Sullivan D, Rohrich RJ. Introducing evidence-based medicine to plastic and reconstructive surgery. Plast Reconstr Surg. 2009; 123:1385–1389. [PubMed: 19337107]
3. Chung KC, Ram AN. Evidence-based medicine: the fourth revolution in American medicine? Plast Reconstr Surg. 2009; 123:389–398. [PubMed: 19116577]
4. Rohrich RJ. So you want to be better: the role of evidence-based medicine in plastic surgery. Plast Reconstr Surg. 2010; 126:1395–1398. [PubMed: 20885263]
5. Burns PB, Chung KC. Developing good clinical questions and finding the best evidence to answer those questions. Plast Reconstr Surg. 2010; 126:613–618. [PubMed: 20679843]
6. Sprague S, McKay P, Thoma A. Study design and hierarchy of evidence for surgical decision making. Clin Plast Surg. 2008; 35:195–205. [PubMed: 18298992]
7. The periodic health examination. Canadian Task Force on the Periodic Health Examination. Can Med Assoc J. 1979; 121:1193–1254. [PubMed: 115569]
8. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989; 95:2S–4S. [PubMed: 2914516]
9. American Society of Plastic Surgeons. Available at: http://www.plasticsurgery.org/Medical_Professionals/Health_Policy_and_Advocacy/Health_Policy_Resources/Evidence-based_GuidelinesPractice_Parameters/Description_and_Development_of_Evidence-based_Practice_Guidelines/ASPS_Evidence_Rating_Scales.html
10. Centre for Evidence Based Medicine. Available at http://www.cebm.net
11. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996; 17:1–12. [PubMed: 8721797]
12. Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the Journal of Bone and Joint Surgery from 1988 through 2000. J Bone Joint Surg Am. 2002; 84-A:388–396. [PubMed: 11886908]
13. Poolman RW, Struijs PA, Krips R, Sierevelt IN, Lutz KH, Bhandari M. Does a "Level I Evidence" rating imply high quality of reporting in orthopaedic randomised controlled trials? BMC Med Res Methodol. 2006; 6:44. [PubMed: 16965628]
14. Urschel JD, Goldsmith CH, Tandan VR, Miller JD. Users' guide to evidence-based surgery: how to use an article evaluating surgical interventions. Evidence-Based Surgery Working Group. Can J Surg. 2001; 44:95–100. [PubMed: 11308245]
15. Thoma A, Farrokhyar F, Bhandari M, Tandan V. Users' guide to the surgical literature. How to assess a randomized controlled trial in surgery. Can J Surg. 2004; 47:200–208. [PubMed: 15264385]
16. Bhandari M, Guyatt GH, Swiontkowski MF. User's guide to the orthopaedic literature: how to use an article about prognosis. J Bone Joint Surg Am. 2001; 83-A:1555–1564. [PubMed: 11679610]
17. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993; 270:2598–2601. [PubMed: 8230645]
18. Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. JAMA. 2000; 284:1290–1296. [PubMed: 10979117]
19. Loiselle F, Mahabir RC, Harrop AR. Levels of evidence in plastic surgery research over 20 years. Plast Reconstr Surg. 2008; 121:207e–211e.

20. McCarthy JE, Chatterjee A, McKelvey TG, Jantzen EM, Kerrigan CL. A detailed analysis of level I evidence (randomized controlled trials and meta-analyses) in five plastic surgery journals to date: 1978 to 2009. Plast Reconstr Surg. 2010; 126:1774–1778. [PubMed: 21042137]
21. Chung KC, Kalliainen LK, Spilson SV, Walters MR, Kim HM. The prevalence of negative studies with inadequate statistical power: an analysis of the plastic surgery literature. Plast Reconstr Surg. 2002; 109:1–6. [PubMed: 11786783]
22. Newman MK, Zemmel NJ, Bandak AZ, Kaplan BJ. Primary breast lymphoma in a patient with silicone breast implants: a case report and review of the literature. J Plast Reconstr Aesthet Surg. 2008; 61:822–825. [PubMed: 17509956]
23. Gaudet G, Friedberg JW, Weng A, Pinkus GS, Freedman AS. Breast lymphoma associated with breast implants: two case-reports and a review of the literature. Leuk Lymphoma. 2002; 43:115–119. [PubMed: 11908714]
24. Sahoo S, Rosen PP, Feddersen RM, Viswanatha DS, Clark DA, Chadburn A. Anaplastic large cell lymphoma arising in a silicone breast implant capsule: a case report and review of the literature. Arch Pathol Lab Med. 2003; 127:e115–e118. [PubMed: 12653596]
25. Keech JA Jr, Creech BJ. Anaplastic T-cell lymphoma in proximity to a saline-filled breast implant. Plast Reconstr Surg. 1997; 100:554–555. [PubMed: 9252643]
26. Duvic M, Moore D, Menter A, Vonderheid EC. Cutaneous T-cell lymphoma in association with silicone breast implants. J Am Acad Dermatol. 1995; 32:939–942. [PubMed: 7751462]
27. Lipworth L, Tarone RE, McLaughlin JK. Breast implants and lymphoma risk: a review of the epidemiologic evidence through 2008. Plast Reconstr Surg. 2009; 123:790–793. [PubMed: 19319041]
28. Lipworth L, Tarone RE, Friis S, et al. Cancer among Scandinavian women with cosmetic breast implants: a pooled long-term follow-up study. International Journal of Cancer. 2009; 124:490–493.
29. Deapen DM, Hirsch EM, Brody GS. Cancer risk among Los Angeles women with cosmetic breast implants. Plast Reconstr Surg. 2007; 119:1987–1992. [PubMed: 17519689]
30. Brisson J, Holowaty EJ, Villeneuve PJ, et al. Cancer incidence in a cohort of Ontario and Quebec women having bilateral breast augmentation. Int J Cancer. 2006; 118:2854–2862. [PubMed: 16381020]
31. Pukkala E, Boice JD Jr, Hovi SL, et al. Incidence of breast and other cancers among Finnish women with cosmetic breast implants, 1970–1999. J Long Term Eff Med Implants. 2002; 12:271–279. [PubMed: 12627789]
32. Brinton LA, Lubin JH, Burich MC, Colton T, Brown SL, Hoover RN. Cancer risk at sites other than the breast following augmentation mammoplasty. Ann Epidemiol. 2001; 11:248–256. [PubMed: 11306343]
33. Denkler K. A comprehensive review of epinephrine in the finger: to do or not to do. Plast Reconstr Surg. 2001; 108:114–124. [PubMed: 11420511]
34. Thomson CJ, Lalonde DH, Denkler KA, Feicht AJ. A critical look at the evidence for and against elective epinephrine use in the finger. Plast Reconstr Surg. 2007; 119:260–266. [PubMed: 17255681]
35. Lalonde D, Bell M, Benoit P, Sparkes G, Denkler K, Chang P. A multicenter prospective study of 3,110 consecutive cases of elective epinephrine use in the fingers and hand: the Dalhousie Project clinical phase. J Hand Surg Am. 2005; 30:1061–1067. [PubMed: 16182068]
36. Chowdhry S, Seidenstricker L, Cooney DS, Hazani R, Wilhelmi BJ. Do not use epinephrine in digital blocks: myth or truth? Part II. A retrospective review of 1111 cases. Plast Reconstr Surg. 2010; 126:2031–2034. [PubMed: 20697319]
37. Wilhelmi BJ, Blackwell SJ, Miller JH, et al. Do not use epinephrine in digital blocks: myth or truth? Plast Reconstr Surg. 2001; 107:393–397. [PubMed: 11214054]
38. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000; 342:1887–1892. [PubMed: 10861325]

Table 1

Canadian Task Force on the Periodic Health Examination's Levels of Evidence*

Level	Type of evidence
I	At least 1 RCT with proper randomization
II.1	Well designed cohort or case-control study
II.2	Time series comparisons or dramatic results from uncontrolled studies
III	Expert opinions

* Adapted from Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J 1979;121:1193-254

Table 2

Levels of Evidence from Sackett*

Level	Type of evidence
I	Large RCTs with clear cut results
II	Small RCTs with unclear results
III	Cohort and case-control studies
IV	Historical cohort or case-control studies
V	Case series, studies with no controls

*Adapted from Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1989;95:2S–4S

Table 3

Levels of Evidence for Prognostic Studies*

Level	Type of evidence
I	High quality prospective cohort study with adequate power or systematic review of these studies
II	Lesser quality prospective cohort, retrospective cohort study, untreated controls from an RCT, or systematic review of these studies
III	Case-control study or systematic review of these studies
IV	Case series
V	Expert opinion; case report or clinical example; or evidence based on physiology, bench research or “first principles”

* Adapted from the American Society of Plastic Surgeons, http://www.plasticsurgery.org/Medical_Professionals/Health_Policy_and_Advocacy/Health_Policy_Resources/Evidence-based_GuidelinesPractice_Parameters/Description_and_Development_of_Evidence-based_Practice_Guidelines/ASPS_Evidence_Rating_Scales.html.

Table 4

Levels of Evidence for Therapeutic Studies*

Level	Type of evidence
1A	Systematic review (with homogeneity) of RCTs
1B	Individual RCT (with narrow confidence intervals)
1C	All or none study
2A	Systematic review (with homogeneity) of cohort studies
2B	Individual Cohort study (including low quality RCT, e.g. <80% follow-up)
2C	“Outcomes” research; Ecological studies
3A	Systematic review (with homogeneity) of case-control studies
3B	Individual Case-control study
4	Case series (and poor quality cohort and case-control study
5	Expert opinion without explicit critical appraisal or based on physiology bench research or “first principles”

*From the Centre for Evidence-Based Medicine, http://www.cebm.net.

Table 5

Grade Practice Recommendations*

Grade	Descriptor	Qualifying Evidence	Implications for Practice
A	Strong recommendation	Level I evidence or consistent findings from multiple studies of levels II, III, or IV	Clinicians should follow a strong recommendation unless a clear and compelling rationale for an alternative approach is present
B	Recommendation	Levels II, III, or IV evidence and findings are generally consistent	Generally, clinicians should follow a recommendation but should remain alert to new information and sensitive to patient preferences
C	Option	Levels II, III, or IV evidence, but findings are inconsistent	Clinicians should be flexible in their decision-making regarding appropriate practice, although they may set bounds on alternatives; patient preference should have a substantial influencing role
D	Option	Level V evidence: little or no systematic empirical evidence	Clinicians should consider all options in their decision making and be alert to new published evidence that clarifies the balance of benefit versus harm; patient preference should have a substantial influencing role

* From the American Society of Plastic Surgeons. Evidence-based clinical practice guidelines. Available at: http://www.plasticsurgery.org/Medical_Professionals/Health_Policy_and_Advocacy/Health_Policy_Resources/Evidence-based_GuidelinesPractice_Parameters/Description_and_Development_of_Evidence-based_Practice_Guidelines/ASPS_Grade_Recommendation_Scale.html. Accessed March 3, 2011

Exhibit 69

J Clin Epidemiol Vol. 51, No. 6, pp. 467–472, 1998

0895-4356/98/$19.00
PII S0895-4356(98)00025-0

COMMENTARY

Observational Research and Evidence-Based Medicine: What Should We Teach Young Physicians?

*Jan P. Vandenbroucke**

Department of Clinical Epidemiology, Leiden University Hospital, Leiden, The Netherlands

There is a continuing uneasiness of proponents of the evidence-based medicine movement with observational research. It pervades the prescriptions for teaching evidence-based medicine, and becomes most acute in the discussion of case-control research. The problem is not new: part of the older "clinical epidemiology" movement had the same problem. In 1978 several controversies resulted in the organization of what was nicknamed the "Bermuda Peace Conference" between opponents and defendants of case-control research, and the proceedings were published as a supplement to the *Journal of Chronic Diseases* [1]. Afterward, the general feeling was that the conference had led to mutual appreciation of the strengths and weaknesses of different forms of clinical and epidemiologic research.

Somehow, the past has come alive again. The continuing misgivings of proponents of evidence-based medicine were most clearly stated in a recent book that aims to teach critical appraisal of the literature to young physicians [2]. In the chapter "Is This Evidence About Harm Important?" ([2], pp. 148, 149) we read: "Because of the biases we described in case-control studies, . . . you might not become impressed with their ROs [risk odds] until they reach 4 or more (some of our colleagues would relax these guides for a serious adverse effect and set them even higher for a trivial one). Since cohort studies are less subject to bias, you might be impressed by RRs [relative risks] of 3 or more in them. And because randomized trials are relatively free of bias, any RR whose confidence interval excludes unity is impressive and warrants further consideration."

Teachers of evidence-based medicine are actively encouraged to propagate brief and simple guidelines. In a paper about teaching critical appraisal, published in the *British Medical Journal*, we read: "Mastering critical appraisal entails learning how to ask a few key questions about the validity of the evidence and its relevance to a particular patient or groups of patients. Its fundamentals can be learnt within a few hours in small tutorials, workshops, interactive lectures, and at the bedside by a wide range of users, including those without a biomedical background" [3].

Crash-course teaching of sweeping generalizations about reading the literature do disservice, not only to the education of the young physician, but also to the intellectual discourse of medicine as a whole, and finally to patients. More than 20 years of advances in principles of observational research, and many worthwhile contributions to etiology and therapeutics, widely ranging from "prone sleeping position" and sudden infant death syndrome [4] to the protection from uterine cancer when coupling progestins to estrogens in hormonal replacement therapy [5], are dismissed without discussion. Most importantly, however, teaching of the role of argument and counterargument by a mix of methodologic and biologic reasoning, which is the hallmark of medical progress, disappears completely.

Criticisms of evidence-based medicine have included the humorous [6], the bitter [7], and the constructive [8]. It is the purpose of this commentary to argue that in teaching young physicians:

1. All study designs in medicine should receive their proper place;
2. The role of argument and counterargument, of which methodologic rules are only a part, should be shown to be equally important in the appraisal of evidence from all types of study design; and
3. We should not be afraid of teaching subjectivity and uncertainty, as these are unescapable.

"CRITICAL APPRAISAL" OF THE LITERATURE?

By emphasizing hard and fast guidelines, evidence-based medicine (EBM) defeats its stated purpose of educating physicians in true "critical appraisal" of the literature. I have witnessed how young physicians who are educated only in EBM become completely lost when they have to think about instances in which randomization is impossible. They do vaguely know that observational designs are quite often the only mode for etiologic and pathophysiologic research, but they do no longer know how to apply the principles, and they see it somehow as "inferior."

*Address for correspondence: Prof. Jan P. Vandenbroucke, MD, PhD, Department of Clinical Epidemiology, Leiden University Hospital, Bldg. 1-C0-P, P.O. Box 9600, 2300 RC Leiden, The Netherlands.

Accepted for publication on 6 March 1998.

An interesting area is that of gene–environment interactions. Clinical researchers who are interested in the influence of genes upon the occurrence of disease have always used observational designs intuitively: after all, the genes were already "randomly" mixed by the patient's parents, and true experimental randomization is impossible. Many of the alleged shortcomings of "retrospective case-control" research play no role when studying genes: the gene does not change upon the occurrence of the disease; possible confounding is usually reduced to a minimum, since people's lifestyles have evolved without any knowledge of them carrying a certain gene, and the influence of other genes can be regarded as "quasi-random." However, clinical and genetic researchers increasingly detect that they should heed proper principles of case-control research. They do no longer collect haphazard cases from their own highly selected practices to compare gene frequencies with that of their laboratory workers and themselves. They do increasingly appreciate that they have to use consecutive unselected cases and appropriate controls. This is especially true in the important and quickly evolving area of interaction between genes and the environment in the causation of disease, which necessitates increasing sophistication in design and execution of case-control research [9]. Sadly, young physicians educated only in EBM are insufficiently equipped to understand case-control reasoning on gene–environment interactions. This whole booming literature will remain a closed book to them, and they will not be able to explain it to their patients: for example when discussing the best mode of contraception with a young woman who has a family history of deep venous thrombosis due to a thrombogenic mutation [10].

In the area of side effects, the EBM guidelines are oblivious to principles of research that are already well established. For example, great emphasis is put on the absolute necessity of the comparability of prognosis between those who use or do not use a certain drug when studying side effects—mimicking the randomization paradigm. This is taught as an absolute requirement and it is written that observational studies, case-control studies in particular, have great difficulty in achieving this requirement. Thereby, the distinction is lost between side effects that are similar to events which may occur in the natural course of the disease, or side effects that are completely unexpected. This distinction has been available from the literature since 1977 [11]. An example was the development of retroperitoneal fibrosis on particular beta-blockers: one need not be preoccupied with any prognostic variable for which the beta-blocker was prescribed to investigate whether beta-blockers give rise to retroperitoneal fibrosis, because the original disease and the development of retroperitoneal fibrosis are unrelated.

Even if the situation is more complex, researchers who study side effects have found good solutions. Two examples from recent controversies about small relative risks are revealing. In the controversy whether the new "third generation oral contraceptives" might cause more venous thrombosis than the older "second generation" preparations, it was said that the original studies were flawed because the users of the newer "third generation products" were more often first-time users, and first-time users have a greater risk (since they have never been challenged by exogenous hormones). A solution came from restricting the comparison between the two classes of contraceptives to first-time users in the first year of use; the two-fold increase in risk was sustained [12]. In another recent controversy about calcium channel blockers (prescribed for hypertension) and the occurrence of cardiovascular disease, it was argued that hypertensives for whom short-acting calcium channel blockers had been prescribed might have a different prognosis. A solution was to restrict the comparison to hypertensives who had been prescribed other "second line" anti-hypertensive drugs; again, the relative risk of 1.6 was sustained [13].

In both instances, the right answer to the objections was not to forget about case-control studies with a relative odds less than a certain arbitrary number. The right answer was to think critically, in true Popperian fashion: to formulate the objections as precise hypotheses and then to look for data that corroborate or negate the hypothesis. The opportunity to teach such distinctions and such critical thinking to young physicians is gone. I have overheard young physicians stating that the controversy on "third versus second generation contraceptives" is unsolvable because "one can never rule out biases in observational studies and a randomized trial is impossible." It is not that these concepts are too difficult: in the above, I have outlined them in a few hundred words.

RANDOMIZATION, IN THEORY . . .

From the teaching of EBM, young physicians recall that randomization is a guarantee for the comparability of two treatment groups. Certainly, randomization should be a guarantee against biased allocation—the tendency of physicians to give the benefit of new treatments to particular patients whom they expect to benefit most. Randomization does not guarantee equality of all prognostic factors, however. Since randomized allocation is "by chance," it is only to be expected that after the randomization the prognosis in the comparison groups might differ—just by chance. That is why tables with "baseline characteristics" are always given in the report of clinical trials. When a baseline difference in prognostic factors occurs, the difference in reasoning between randomized and nonrandomized studies disappears completely. The oldest, and perhaps still the nicest, teaching example was the University Group Diabetes Program (UGDP) study on the control of diabetes [14]. In that trial one of the treatment groups (on a promising drug) showed more cardiovascular mortality. Critics were quick to point out that "just by chance" that group also had a somewhat

higher baseline cardiovascular risk. In a beautiful rebuttal, Cornfield showed by logical argument and stratification on risk factors that the perceived difference in baseline prognosis could never account for the higher incidence in the treatment group [15]. This way of arguing, however, is exactly the same as with observational research: it rests upon prior knowledge, judgment, and stratification. Inversely, it becomes clear that the "acceptance." of the equality of two treatment groups, when randomization has resulted in little or no difference at baseline, is an equally subjective judgment. It is not surprising that the same Cornfield, who gave us both the odds ratio and the logistic model, later wrote that his (admittedly Bayesian) views placed ". . . emphasis on reasonable scientific judgement and accumulation of evidence and not on dogmatic insistence on the unique validity of a particular procedure" [16].

. . . AND IN PRACTICE

What happens to randomized trials in actual practice? An interesting review showed that over a large number of published trials, the treatment group was repeatedly and statistically significantly smaller than the placebo group [17]. Apparently, patients are slightly more often removed from the treatment group—despite lip service to "blinding" and "intention to treat analysis." It has also been shown that randomized trials with outside sponsors report more often statistically significant advances [18]. How can we explain this?

Somehow, randomized controlled trials do show a "willingness to please." In a meta-analysis on low molecular weight heparin prophylaxis for deep venous thrombosis it was demonstrated that the trials with a slightly lesser quality score (always somewhat arbitrary, of course) showed the expected effects of these agents: they showed "less thrombosis as well as less bleeding" [19]. Trials with higher quality scores, generally published in journals with higher impact factors, also showed less thrombosis, but, in contrast, somewhat more bleeding with the newest heparin [19]. These meta-analytic observations were on a number of smaller trials: overall, they favored the newer heparins both in being more effective and in having fewer side effects, which to some observers was the final message. A large trial was published after the meta-analysis, however, and found no difference in the risk of thrombosis but quite a dramatic reduction in bleeding [20]. This discrepancy was used as evidence in an overview on the conflicting results between meta-analyses and single large trials [20].

Which should we believe: the large trial, the overall meta-analysis result, or the subgroup of higher quality trials? To understand what happened, it is highly interesting to note that the original biochemical insights in the mode of action of low molecular weight heparins had indeed raised the hope that they would be at least as effective as anticoagulants and at the same time show less bleeding—later biochemical studies, however, showed that a therapeutic benefit over classic heparin was not that likely (apart of course from the easier mode of administration of the newer heparins, which is an undisputed benefit) [21]. Apparently, the interpretation of many of the earlier smaller trials was somehow guided by the original biochemical insights, while later authors were more equipoised.

Neither randomization nor blinding are guarantees against "willingness to please," which, I hasten to add, is a complex phenomenon that can result from the best and honest intentions. For example, I suspect that an all too human "willingness to please" might be part of the explanation why results of randomized controlled trials on homeopathy show that these "infinite dilutions" work. Meta-analyses exist on "good quality randomized trials" showing that homeopathy might be of value for a wide range of conditions [22]. The problem with homeopathy is not that no mechanism for its action is known; the problem is that any pharmacologic activity of an "infinite dilution" is impossible. Since a randomized trial of "infinite dilutions" versus "solution only" is a game of chance between two placebos, positive trials of homeopathy can only result from chance. Given the large number of positive homeopathy trials, however, and a sensitivity analysis about the "file drawer problem" in the most recent meta-analysis [22], it is highly unlikely that selective publication only can account for the several positive meta-analyses. Other subjective factors must be at play. Before we dismiss randomized trials on homeopathy, however, we should not forget that the same mechanisms might explain some positive results of randomized trials in regular "scientific" medicine.

There is another practical side to randomized trials. Many randomized controlled trials on new drug developments are supported by outside sponsors. At worst—a practice that is hopefully declining, as it is even seen as undesirable by responsible sponsors—the trial is seeded with some important medical names and carried out by a specialized bureau that offers trial nurses, statisticians, and a first draft of the paper by professional writers, who have already casted it in the mode of the intended journal. While such publications might find their way more easily in smaller journals or their supplements, larger trials are not completely immune.

Recently it has become clear that outside sponsors still reserve the contractual right to stop large trials without having to explain why [23,24]. In each instance where it happens, one can only second-guess the reasons. They lead to suspicions that are in the end not in the interest of the sponsors themselves. Carrying out large trials has become big enterprise. The principal investigator's academic departments become very dependent on a huge cash flow, also for truly interesting side-studies, and one is not surprised that there is always the interesting subgroup analysis or recombination of endpoints (or both) that might be reported as tentatively positive, and become the start of a new trial. A deplorable side effect of the largeness of the revenues of

directly sponsored trials is that practising physicians tend to become less attracted to trials that do not have such revenues. The realities of life in the world of randomized trials contrast starkly with the ethereal views of the non-biasedness of the randomization procedure. Often I have more sympathy with observational research, even in small journals, where the only error committed might be one of youthful enthusiasm or of ignorance; when pointed out to the young authors, both are easily corrected without grudge.

A MIXTURE OF EVIDENCE . . .

Proponents of evidence-based medicine are known to be among the most sincere advocates of the patient's right to be diagnosed and treated by the best of scientific knowledge. Archie Cochrane was the great champion of this idea [25]. However, when making medicine singularly dependent on randomized trials, of which some unverifiable part is of the types that I just described, I am not certain that the best interest of the patient is always served. Moreover, by taking away credibility from case-control research, which is the most potent research design to study side effects, the patient and the physician are left defenseless a second time. Outside parties that do not like a particular case-control result on a side effect will quickly reach to the teachings of EBM to tell young physicians that all such results are nonsense. They already do.

Judging observational studies appears to be more difficult than judging randomized trials, because it involves a process of "argument and counter-argument" for which there are general principles only, but which is different in each instance. It cannot be done by a few rules of thumb. A good example of this argumentational process has been described in an attempt to decide from observational epidemiologic studies whether alcohol really protects against cardiovascular disease [26]; the side effect examples cited above show the same pattern. However, the same argumentational process is inherent in randomized trials when one judges the comparability of baseline prognostic factors in the treatment groups, the appropriateness of the dosage and timing of the drug, the exclusion and inclusion criteria, or the definition of the endpoints.

. . . ALSO IN META-ANALYSIS

In recent discussions about the merits and shortcomings of meta-analysis, it has become clear that meta-analysis is itself a kind of "observational research," also nicknamed "the epidemiology of results"—even if it is a meta-analysis of randomized trials. New insights lead to the notion that the appraisal of a meta-analysis also rests on a process of arguments and counterarguments, that is based on a "mixture of evidence." Egger *et al.* showed how a large part of published meta-analyses have asymmetrical funnel plots [27]. The classic explanation for such skewed funnel plots was "publication bias." Egger *et al.*, however, convincingly argue that a host of other factors can be responsible. All kinds of distortions can lead to asymmetrical funnels: data massage in the analysis, poor blinding during data collection, unclear stopping rules, or downright fraud. Alternatively, it can also be true underlying heterogeneity, or the "wrong parametrisation" (funnels that are asymmetrical on a relative risk scale might not be on a risk difference scale, and vice versa). It can also be a mixture of all these things. Very often, we do not care: colleagues who are involved in clinical trials have told me that they suspect that "on average" all clinical trials are slightly too optimistic in their results—they only indicate the direction and the order of magnitude. This is fine as most practicing physicians know that the literature is always somewhat optimistic and that after initial waves of enthusiasm about the newest intervention, there always follows a wave of disenchantment. Perhaps, we should always take the whole literature—on whatever subject—with a grain of salt.

The biggest problem, however, is what to do with a meta-analysis of randomized trials that show "skewed funneling" when there is disagreement and controversy about the interpretation [20,27]. Does one simply state that the whole thing is "biased," and dismiss all randomized trials on the subject? Then, one dismisses the results of one group of investigators because of the results of a completely different group. Sometimes "quality criteria" or criteria about study size will help, but the example about homeopathy teaches us that these do not always settle the issue. The low molecular weight heparin example teaches us that much is gained when going back to basic biology. Thus, there seems no escape from a "return to the expert" who tells us which trial he does believe most, based on his insights in pathophysiology, pharmacology, and, of course, also some methodology. The only guarantee that we can ask from the expert is that he or she states his reasoning as clearly as possible.

At certain points we should live with the fact that the experts cannot agree whether a particular intervention in a particular pivotal trial was administered in the right way: right dosage, timing, right patients, etc. We might even live with the fact that in certain areas of heated controversy, where "statistical significance" is barely reached by a mix of trials of different origin, quality and size, some experts suspect that enough trials might be tainted by some "willingness to please" (even if the funnel is not downrightly skewed) as to withhold judgment. That withholding of judgment becomes purely subjective: it is because we know about the stakes in a particular controversy and about the honest, but self-delusive potential of the human brain. We might even teach this to our students and young physicians—as such has always been the nature of medicine.

None of this implies that one should always give precedence to "basic science" arguments over numerical arguments. When they are strong, or preclude any mechanism

of action, like with homeopathy, one tends to put faith on the physico-chemical side of medical reasoning. When biological arguments are largely absent, one goes with numerical evidence, especially if it has undergone several cycles of argument and counterargument. Most often, there is enough information to rely on a mixture of both. The relative input of the different sources of medical knowledge in that mixture will always remain a subjective and continually changing choice.

THE FUTURE EDUCATION OF YOUNG PHYSICIANS

The December 1997 issue of the *British Medical Journal* contained a heartening Christmas message by leaders of the evidence-based medicine movement: "Choosing the best research design for each question" [28]. It called for a moratorium on discussions on which research design is best—and thereby has the spirit of the "Bermuda Peace Conference" of two decades ago: "Each method should flourish, because each has features that overcome the limitations of the others when confronted with questions they cannot reliably answer" [28].

In the future, young physicians might be taught that study designs, like all drugs in medicine, have their indications and contraindications, and that there is no hierarchy. They should be taught for what issue in what circumstance a particular design might be most appropriate, and also what the shortcomings are of each. For example, besides the advantages, the shortcomings of prospective follow-up studies (low number of endpoints and the great temporal distance between initial exposure and outcome) should be explained—to be on their guard where appropriate. The role of case-series should be highlighted: for the description of new diseases, for studying pathogenesis, or for attracting attention to strange observations that might be of future scientific or educational value. Perhaps we should not shrink away from calling methodology itself into question: we might, at the end of a course, give examples in which one dismisses a randomized controlled trial right away, even if on the surface one cannot find anything wrong (e.g., randomized trials that show no strong mortality benefit from smoking cessation), and inversely, how a single case observation led the way to new therapeutic insights. We might teach them historical examples of side effect brawls with different outcomes: not only the breast cancer and reserpine example, but also the aspirin and Reye's syndrome story in which epidemiology was first vilified and brought to court, but later proved triumphant. We can explain to our students the role of rhetoric, for example in the way in which numerical results are explained: as percentages, relative odds, risk differences of numbers-needed-to-treat. Perhaps we might also discuss "conflicts of interest" and the different approaches and attitudes that are taken toward it, both philosophically [29] and in daily reality [30].

The education of young physicians should overcome the level of the few key questions in a crash course. Of course, they should not all become methodologic or theoretical experts. They will in the first place be practitioners, but they should become intelligent consumers of literature and especially of arguments. This does not mean that each of them should individually judge whatever piece of the literature on whatever topic: that is impossible since that demands also subject matter knowledge. But they should at least be familiar with the terms of the debate, the way of arguing, so that they can recognize what arguments are used in what circumstances, and whether they want to be swayed by them. In the end, they should recall that "The art is long, life short," meaning in this instance that medical history has shown repeatedly that erring is possible, despite the best of arguments. We should not be afraid of teaching them uncertainty.

References

1. The case-control study: Consensus and controversy. **J Chronic Dis** 1979; 32: 1–144.
2. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. **Evidence-Based Medicine: How to Practice and Teach EBM**. New York, NY: Churchill Livingstone; 1997.
3. Rosenberg W, Donnald A. Evidence based medicine: An approach to clinical problem-solving. **BMJ** 1995; 310: 1122–1126.
4. de Jonge GA, Engelberts AC, Koomen-Liefting AJ, Kostense PJ. Cot death and prone sleeping position in The Netherlands. **BMJ** 1989; 298: 722.
5. Voigt LF, Weiss NS, Chu J, Daling JR, McKnight B, van Belle G. Progestagen supplementation of exogenous oestrogens and risk of endometrial cancer. **Lancet** 1991; 338: 274–277.
6. Grahame-Smith D. Evidence based medicine: Socratic dissent. **BMJ** 1995; 310: 1126–1127.
7. Shahar E. Re: Evidence based medicine. **Lancet** 1996; 347: 1172.
8. Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence-based medicine." **Am J Med** 1997; 103: 529–535.
9. Khoury MJ. From genes to public health: The applications of genetic technology in disease prevention. **Am J Public Health** 1996; 86: 1717–1722.
10. Vandenbroucke JP, van der Meer FJ, Helmerhorst FM, Rosendaal FR. Factor V Leiden: Should we screen oral contraceptive users and pregnant women? **BMJ** 1996; 313: 1127–1130.
11. Jick H. The discovery of drug-induced illness. **N Engl J Med** 1977; 296: 481–485.
12. Poulter NR, Farley TMM, Chang CCL, Marmot MG, Meirik O. Authors' reply: Safety of combined oral contraceptive pills. **Lancet** 1996; 347: 547.
13. Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Raghunathan TE, Weiss NS, *et al*. The risk of myocardial infarction associated with antihypertensive drug therapies. **JAMA** 1995; 274: 620–625.
14. Marks HM. The progress of experiment: Science and therapeutic reform in the United States, 1900–1990. Cambridge, UK: University Press; 1997: 197–228.
15. Cornfield J. The University Group Diabetes Program; a further statistical analysis of the mortality findings. **JAMA** 1971; 217: 1676–1687.
16. Cornfield J. Recent methodological contributions to clinical trials. **Am J Epidemiol** 1976; 104: 408–421.

17. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. **Lancet** 1990; 335: 149–153.
18. Davidson RA. Source of funding and outcome of clinical trials. **J Gen Int Med** 1986; 1: 155–158.
19. Nurmohamed MT, Rosendaal FR, Büller HR, Dekker E, Hommes DW, Vandenbroucke JP, Briët E. Low molecular weight heparin versus standard heparin in general and orthopedic surgery: A meta-analysis. **Lancet** 1992; 340: 152–156.
20. LeLorier J, Gregoire G, Genhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized controlled trials. **N Engl J Med** 1997; 337: 536–542.
21. Thomas DP. Does low molecular weight heparin cause less bleeding? **Throm Haemost** 1997; 78: 1422–1425.
22. Linde K, Clausius N, Ramirez G, Melchart D, Eitel F, Hedges LV, Jonas WB. Are the clinical effects of homeopathy placebo effects? A meta-analysis of placebo-controlled trials. **Lancet** 1997; 350: 834–843.
23. Editorial. Good manners for the pharmaceutical industry. **Lancet** 1997; 349: 1635.
24. Editorial. A curious stopping rule from Hoechst Marion Roussel. **Lancet** 1997; 350: 155.
25. Cochrane AL. **Effectiveness and efficiency: Random reflections on health services**. Nuffield, UK: Nuffield Provincial Hospital Trust; 1972.
26. Maclure M. Demonstration of deductive meta-analysis: Ethanol intake and risk of myocardial infarction. **Epidemiol Rev** 1993; 15: 328–351.
27. Egger M, Smith GD, Scheider M, Minder C. Bias in meta-analysis detected by a simple graphical test. **BMJ** 1997; 315: 629–634.
28. Sackett DL, Wennberg JE. Choosing the best research design for each question. **BMJ** 1997; 315: 1636.
29. Rothman KJ, Cann CI. Judging words rather than authors. **Epidemiology** 1997; 8: 223–225.
30. Stelfox HT, Chua G, O'Rourke K, Detsky AS. Conflict of interest in the debate over calcium channel antagonists. **N Engl J Med** 1998; 338: 101–106.

Exhibit 70

Interpretation of Epidemiologic Studies on Talc and Ovarian Cancer

prepared by

Kenneth J. Rothman
Harris Pastides
Jonathan Samet

November 28, 2000

Executive Summary

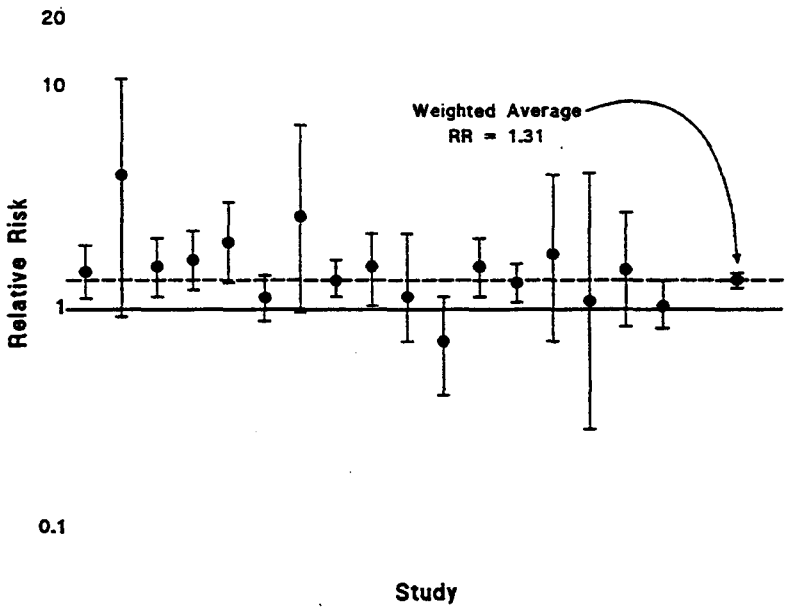
A weighted average of the results from epidemiologic studies to date measuring the relation between talc and ovarian cancer risk gives an overall relative risk of 1.31, with a 95% confidence interval of 1.21–1.41. Bias and causation are competing explanations for the weak positive association observed. This weak association could be an underestimate of a stronger association if there are errors in measuring talc exposure that apply uniformly to all study subjects (nondifferential misclassification). On the other hand, nondifferential misclassification does not bias an association that is null to begin with, so postulating nondifferential misclassification cannot shed light on whether the association results from a causal relation or not. Most of the published studies are interview-based case-control studies, subject to recall bias, which can readily give rise to associations of this magnitude. The evidence from these studies regarding recall bias is mixed. Uncontrolled confounding can also easily explain associations this weak; although no single confounding factor would seem to account for the overall effect, the combined effect of several such unidentified confounders could do so. In considering these competing explanations of bias and causation, the evidence in favor of a causal explanation is only the overall weak association of a relative risk of 1.31. The lack of a plausible biologic mechanism, on the other hand, weighs against a causal interpretation. Also weighing against a causal explanation is the dose-response pattern among talc users, which is an inverse trend for both duration of use and frequency of use. A causal relation would predict a positive trend, not an inverse trend. Based on these considerations, we suggest that the evidence to date does not indicate that talc can be "reasonably anticipated to be a human carcinogen."

Introduction

In this document we offer an interpretation of the epidemiologic literature with respect to the causal hypothesis that talc exposure causes an increase in the occurrence of ovarian cancer. Overall, we identified 23 epidemiologic studies conducted since 1980 that have examined consumer talc exposure with respect to subsequent risk for ovarian cancer.[1-23] The search methodology is described in the appendix. Sixteen of these were case-control studies reporting new data with effect estimates for talc exposure,[2-5,7,10,11,13-15,17-19,21-23] and one was a cohort study reporting an effect estimate.[9] One study examined occupational exposure to talc in women, but there were few exposed women in this study[16]; the other studies did not report quantitative effect estimates. The importance of this comparatively small set of epidemiologic studies is underscored by the paucity of relevant animal research on this question.

Most of these published reports come from epidemiologic studies in which talc was not the primary focus. Perhaps for this reason, talc exposure information was often crude. In only a few of these studies was there any attempt to categorize talc exposure by frequency of use or duration of use. For the 17 studies that reported some epidemiologic measure of effect, it was usually a relative risk estimate for ovarian cancer given that there was some exposure to talc, compared with no exposure or minimal exposure. These results are depicted graphically in figure 1. The findings on balance indicate a slight positive association between talc exposure and ovarian cancer, with an overall weighted relative risk of 1.31, and a 95% confidence interval of 1.21–1.41.

Figure 1
Study-specific Relative Risk Estimates for Ovarian Cancer Among Talc Users, and Overall Weighted Average of Study Results



Issues Affecting Causal Inference

Inferring a causal relation from a pattern of epidemiologic results follows no recipe, but certain principles can be applied. To begin with, what alternative explanations might be offered to explain a pattern of positive findings? If an uncontrolled confounding factor or a study-related bias could explain the results, a causal inference is less reasonable. Second, is there a plausible biologic mechanism? For example, environmental tobacco smoke shows a weak association with lung cancer in numerous epidemiologic studies of never smokers, but the plausibility of the relation, based on the known constituents of the smoke and their effect in higher concentrations, among active smokers, makes a causal inference more reasonable. Third, is there a consistent dose-response trend in the data? With rare exception, every causal relation in epidemiologic research shows a progressive relation between various measures of increasing exposure. In this discussion paper, we address the following issues that we believe are potentially relevant to causal inference regarding talc and ovarian cancer:

1. Exposure misclassification
2. Recall bias
3. Confounding
4. Dose-response trends
5. Biologic mechanism

Below we discuss briefly the import of each of these topics with respect to the interpretation of the epidemiologic literature of talc and ovarian cancer. We omit discussion of the role of chance in explaining any of the findings, because the combined weight of the 17 studies in figure 1 indicates that chance alone is an unlikely explanation for the overall weighted average of relative risks from the studies of 1.31. Other possible issues, such as selection biases and reverse causation might be relevant, but appear less important to us in interpreting these results, so we have omitted them in the interests of brevity. (Reverse causation, for example, could occur if preclinical ovarian cancer prompted women to use talc; while this situation is possible in some instances, we do not think it is a realistic explanation for the observed effects.)

Exposure Misclassification

Nearly all the studies were case-control studies. It is commonly believed that the validity of case-control studies is worse than that of cohort studies, but this view is mistaken. The validity of a study depends on the specifics of the study design, the nature of the data, and the nature of the hypothesis that the study addresses. For example, a cohort study that examines the long-term risk of cancer among coffee drinkers after a one-time dietary assessment of coffee consumption would suffer from weak exposure assessment. Although the exposure information might be accurate for the time at which it was collected, the exposure status of cohort members will change with time and the initial measure might be only poorly correlated with a more meaningful measure of coffee consumption. The effect of having a poor measure of exposure will be considerable nondifferential misclassification, a type of error that introduces a bias into study results that tends to drive effect estimates towards the null condition of no effect. In contrast, it may be possible to get more detailed exposure information from study subjects in a case-control study, which might thus avoid some of the bias that would result from a cohort study.

Much like coffee consumption, talc exposure is likely to vary over time as women age and their reasons for deciding to use talc change. Consequently a single baseline assessment of talc exposure at the start of follow-up in a cohort may lead to effect estimates that are biased toward the null. If talc habits are steady over time, a single baseline assessment becomes more informative. Furthermore, if talc use influences cancer risk with a long induction period, talc assessment at the start of a cohort study is more meaningful than an assessment of coffee drinking on heart disease risk, which is thought to have only a short-term effect.

Case-control studies also suffer from exposure misclassification, but the potential exists to extract more detailed history of exposure from the subject interview. In most of these studies, the exposure metric is based on interview information. It is subject to inaccuracies from recall error, as well as inaccuracies reflecting the nature of the questions asked and their relation to any biologically relevant measure of talc exposure. Ideally one would wish to have a measure of talc dose within the upper reproductive tract. The actual measures obtained by interview, however, are likely to be only modestly correlated with a hypothetically ideal measure. The result of this inevitable non-differential misclassification would be to bias any real effect towards the null. Nevertheless, one cannot draw the conclusion that the overall slight positive relation between talc exposure and ovarian cancer must be an underestimate of a larger effect because of nondifferential misclassification. Non-differential misclassification does not introduce any bias toward the null if the association is null to begin with, so to draw the conclusion that the overall effect estimate from the 17 studies is an underestimate, one must already know or assume that there is an even stronger positive relation in the data. Thus, the prospect of non-differential misclassification in measuring talc exposure does not provide any help by itself in assessing whether talc is related to ovarian cancer.

Recall Bias

Cohort studies do not suffer from recall bias, but recall bias is an issue for case-control studies that obtain exposure information from subject interviews. Such was the case for all the case-control studies whose effects are summarized in figure 1. Recall bias can readily introduce enough bias to produce the modestly-sized overall effect (RR = 1.3) that emerges from these studies. As an example, one of us reported an association between Bendectin and congenital heart disease in 1979, with a RR of 1.6.[24] One possibility for that positive relation was recall bias, a strong consideration in light of the study design that produced the finding (the study was not designed to evaluate Bendectin, which was only an incidental finding). To resolve the issue, a second study was undertaken, this time aimed at evaluating an effect of Bendectin by eliminating recall bias using a different design.[25] The second study found a RR of 1.0, prompting the conclusion that the RR of 1.6 reported in the earlier study was due to recall bias. The amount of recall bias for Bendectin in the 1979 study amounted to an apparent effect that was much stronger than the overall effect estimate for talc and ovarian cancer in the combined studies in figure 1.

We believe that there is mixed evidence for recall bias in these studies. We base this interpretation on the few studies that examined the effect of talc separately among women who had a tubal ligation and those who did not. If recall bias were the explanation for the full effect seen in the published literature, we would predict that the effect of talc exposure would appear to be about the same for women who have a tubal ligation and those who did not, because tubal ligation is unlikely to affect recall bias. In contrast, it would likely affect any biologic action of

talc. Only three studies give information relevant to this question. In those studies, the evidence is mixed. In one study the effect of talc is greater among women who have not had a tubal ligation,[22] and in a second, talc use appeared to have no adverse effect among women who had either a hysterectomy or a tubal ligation.[23] In the third study,[2] however, there was little difference in the effect of talc for women with and without tubal ligation or hysterectomy and the effect for both groups was near null. Thus, the overall evidence on the possibility of recall bias is equivocal, with no clear answer as to whether recall bias can be eliminated as an explanation.

Confounding

Although there are some strong risk factors for ovarian cancer, for any of them to be confounding to an extent that could account for the positive relations that have been reported, they would have to be strongly correlated with talc use. Family history, ethnicity, obesity and some reproductive risk factors are positively associated with the risk of ovarian cancer, but the magnitude of these associations does not appear high enough to introduce enough confounding, even jointly, to explain completely the positive association. Of course, it remains possible that yet unidentified risk factors for ovarian cancer could be important confounders, and several such factors in the aggregate could give risk to an overall association as weak as the one between talc and ovarian cancer.

Dose-response trends

A nearly constant feature of causal relations in epidemiology and in the pathogenesis of cancer in particular is a monotonically increasing relation between measures of exposure and disease risk. Even when disease risk increases through a threshold phenomenon, progressive dose-response trends are observed because the exposure measure varies and smooths the step relation of a threshold into a gradual climb in risk. In contrast, many biases would not produce a monotonic dose-response relation. For example, Horwitz and Feinstein advanced a theory of "detection-bias" as a non-causal alternative to the theory that exogenous estrogens cause endometrial cancer.[26] According to this theory, administration of estrogens would provoke genital bleeding among some women, leading to a work up and to the diagnosis of pre-existing endometrial cancers, accounting for the observed association. This theory, however, predicted that the increase in endometrial cancer risk would be greatest for short-term users of exogenous estrogens and would decline toward no effect for longer-term users. In actual fact this inverse dose-response trend was not observed, undermining the detection bias theory.

Exposure to talc can be characterized by the age at which use started, the number of years of use, and the frequency of use (e.g., number of times per day or per week). Among the talc studies, several reported on either frequency of talc use or duration of talc use, or both. We combined the findings from these studies into a meta-regression,[27] an analysis that combines dose-specific information from various studies into a single weighted regression analysis. Each data point in a meta-regression represents one effect estimate at a given dose level; the data points are weighted by the precision of each estimate, back-calculated from the confidence interval for that estimate.

In figure 2 we show the data points and meta-regression line for frequency of talc use, and in figure 3 for duration of talc use. These regression analyses confirm the picture that one obtains from reading the individual studies (table 1): the dose-response relation across dose levels above zero for talc exposure is not increasing, but instead declines. Although

misclassification could flatten a dose-response curve, it would not produce an inverse dose-response curve. Thus, the observed pattern, whether based on individual studies or from the combined meta-regression analysis, is not consistent with a causal interpretation for talc exposure. Instead it suggests that some as yet unidentified bias accounts for the overall modest relation between talc exposure and ovarian cancer.

Figure 2
Trend in Relative Risk by Frequency of Talc Use Among Users

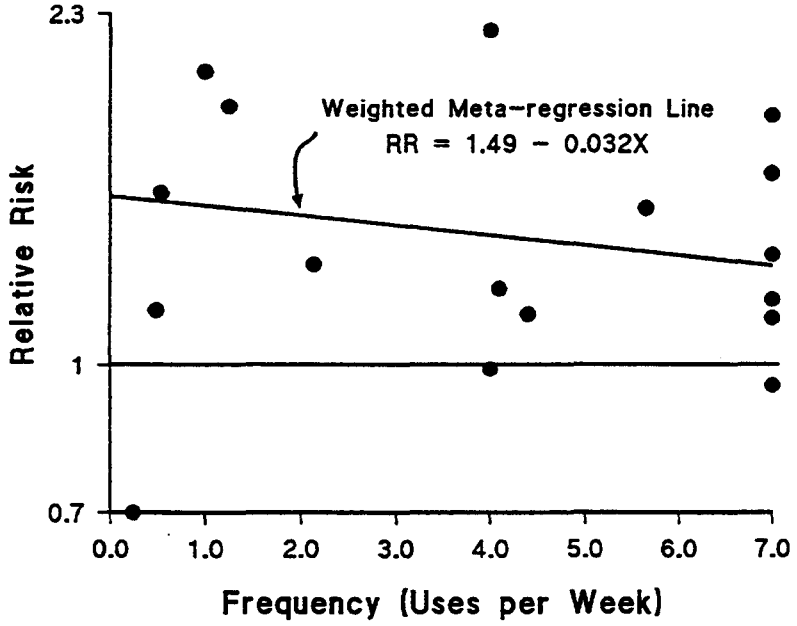


Figure 3
Trend in Relative Risk by Duration of Talc Use Among Users

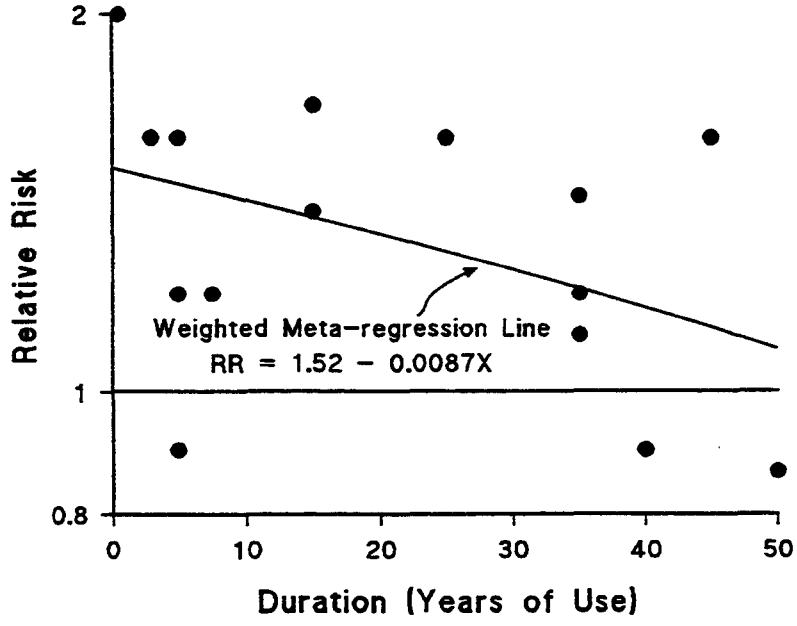


Table 1
Relative Risk Estimates of Ovarian Cancer by Frequency and Duration of Talc Use*

Citation	Frequency (Applications/wk)	Relative Risk	95% Confidence Interval
Booth et al.1989	7.00	1.30	0.80-1.90
	1.00	2.00	1.30-3.40
	0.25	0.70	0.30-1.80
Chang and Risch 1997	1.25	2.00	1.24-2.73
	4.40	1.13	0.74-1.72
	7.00	0.95	0.61-1.49
Cramer et al. 1999	4.00	2.21	1.37-3.56
	7.00	1.17	0.78-1.76
	7.00	1.57	0.80-3.10
Gertig et al. 2000	0.50	1.14	0.81-1.59
	4.00	0.99	0.67-1.46
	7.00	1.12	0.82-1.55
Harlow et al. 1992	0.55	1.50	0.80-2.70
	4.10	1.20	0.60-2.20
	7.00	1.80	1.10-3.00
Whittemore et al. 1988	2.14	1.27	0.82-1.96
	5.65	1.45	0.94-2.22

Citation	Duration (years)	Relative Risk	95% Confidence Interval
Chang and Risch 1997	15	1.70	1.09-2.64
	35	1.44	0.96-2.15
	50	0.86	0.54-1.38
Harlow et al. 1992	5	1.20	0.50-2.60
	25	1.60	1.00-2.70
	45	1.60	1.00-2.70
Ness et al. 2000	1	2.00	1.00-4.00
	3	1.60	1.10-2.30
	7.5	1.20	0.80-1.90
	35	1.20	1.00-1.50
Whittemore et al. 1988	5	1.60	1.00-2.57
	35	1.11	0.74-1.65
Wong et al. 1999	5	0.90	0.60-1.50
	15	1.40	0.90-2.20
	40	0.90	0.60-1.20

* For Open-ended Categories, the Values Assigned Assume that the Upper Category Boundary Corresponds to a Maximum Frequency Equal to Daily Use and a Maximum Duration of Use of 60 Years

Biologic Mechanism

The most plausible biological mechanism relating to the development of ovarian cancer concerns ovulation and the hormonal factors affecting it. Specifically, factors that suppress ovulation, such as gravidity, breast feeding, oral contraceptive use, tubal ligation and hysterectomy appear to reduce strongly the risk of ovarian cancer. Body mass index may also affect ovarian cancer risk. Medical conditions that may affect ovulation and also appear to increase the risk of ovarian cancer include endometriosis, ovarian cysts, and hyperthyroidism.

It does not appear plausible, however, that talc exposure has a direct effect on ovulation. If talc exposure is correlated with factors that affect ovulation, that correlation would produce confounding, as discussed above. If talc were a cause of ovarian cancer, it is presumably through a different mechanism than the many risk factors already known to affect ovarian cancer risk. There is no other evidence regarding such a mechanism, nor any clear evidence that talc applied perineally or on diaphragms makes its way physically to the ovaries. Ness et al suggest that inflammation may mediate ovarian cancer risk and that talc may play a role by causing inflammation.[17] This theory merits further investigation, although the tenability of the theory rests on the issue of whether talc particles physically reach the ovaries. Without a clear biologic mechanism for talc to cause ovarian cancer, an inference that talc does cause ovarian cancer would be an example of a "black-box" inference, meaning that the inference lacks a biologic foundation. "Black-box" inferences, such as the inference some draw that electromagnetic fields increase the risk for various cancers, are not necessarily invalid, but they are inherently more tenuous than inferences that are rooted in biologic explanations.

Conclusion

The only evidence to support a causal interpretation is the overall modest positive association seen in most of the epidemiologic studies that we have cited. The association is weak enough to be plausibly explained by unidentified bias. Recall bias is one possibility, but unidentified confounding could also readily give rise to the weak level of association that confronts us from these studies. Bias and causation are competing explanations for the weak positive association observed, an association that could be an underestimate of a stronger real association if nondifferential misclassification has diluted it. In considering these competing explanations, the lack of a plausible biologic mechanism based on the evidence to date weighs against a causal interpretation. More important, there is also positive evidence against a causal association: the inverse dose-response trend for both duration of use and frequency of use, a pattern that could not be explained by a causal relation. Based on these considerations, we suggest that the evidence to date does not indicate that talc can be "reasonably anticipated to be a human carcinogen."

References

1. Booth M, Beral V, Smith P: Risk factors for ovarian cancer: A case-control study. Br J Cancer 1989;60:592-598.

2. Chang S, Risch HA: Perineal talc exposure and risk of ovarian carcinoma. Cancer 1997;79:2396-2401.

3. Chen Y, Wu PC, Lang JH, Ge WJ, Hartge P, Brinton LA: Risk factors for epithelial ovarian cancer in Beijing, China. Int J Epidemiol 1992;21:23-29.

4. Cook LS, Kamb ML, Weiss NS: Perineal powder exposure and the risk of ovarian cancer. Am J Epidemiol 1997:145:459-465.

5. Cramer DW, Liberman RF, Titus-Ernstoff L, Welch WR, Greenberg ER, Baron JA, Harlow BL: Genital talc exposure and risk of ovarian cancer. Int J Cancer 1999; 81:351-356.

6. Cramer DW, Xu H: Epidemiologic evidence for uterine growth factors in the pathogenesis of ovarian cancer. Ann Epidemiol 1995;5:310-314.

7. Cramer DW, Welch WR, Scully RE, Wojciechowski CA: Ovarian cancer and talc. Cancer 1982;50:372-376.

8. Eltabbakh GH, Piver MS, Natarajan N, Mettlin CJ: Epidemiologic differences between women with extraovarian primary peritoneal carcinoma and women with epithelial ovarian cancer. Obstet Gynecol 1998;91:254-259.

9. Gertig DM, Hunter DJ, Cramer DW, Colditz GA, Speizer FE, Willett WC, Hankinson SE: Prospective study of talc use and ovarian cancer. J Natl Cancer Inst 2000;92:249-252.

10. Godard B, Foulkes WD, Provencher D, Brunet JS, Tonin PN, Mes-Masson AM, Narod SA, Ghadirian P: Risk factors for familial and sporadic ovarian cancer among French Canadians: A case-control study. Am J Obstet Gynecol 1998;179:403-410.

11. Green A, Purdie D, Bain C, Siskind V, Russell P, Quinn M, Ward B, and the Survey of Women's Health Study Group: Tubal sterilisation, hysterectomy and decreased risk of ovarian cancer. Int J Cancer 1997;71:948-951.

12. Hankinson SE, Hunter DJ, Colditz GA, Willett WC, Stampfer MJ, Rosner B, Hennekens CH, Speizer FE: Tubal ligation, hysterectomy, and risk of ovarian cancer. A prospective study. JAMA 1993;270:2813-2818.

13. Harlow BL, Cramer DW, Bell DA, Welch WR: Perineal exposure to talc and ovarian cancer risk. Obstet Gynecol 1992; 80:19-26.

14. Harlow BL, Weiss NS: A case-control study of borderline ovarian tumors: The influence of perineal exposure to talc. Am J Epidemiol 1989;130:390-394.

15. Hartge P, Hoover R, Lesher LP, McGowan L: Talc and ovarian cancer. JAMA 1983;250:1844.

16. Hartge P, Stewart P: Occupational and ovarian cancer: A case-control study in the Washington, DC, metropolitan area, 1978-1981. J Occup Med 1994;36:924-927.

17. Ness RB, Grisso JA, Cottreau C, Klapper J, Vergona R, Wheeler JE, Morgan M., Schlesselman JJ: Factors related to inflammation of the ovarian epithelium and risk of ovarian cancer. Epidemiology 2000;11:111-117.

18. Purdie D, Green A, Bain C, Siskind V, Ward B, Hacker N, Quinn M, Wright G, Russell P, Susil B: Reproductive and other factors and risk of epithelial cancer: An Australian case-control study. Survey of Women's Health Study Group. Int J Cancer 1995;62:678-684.

19. Rosenblatt KA, Szklo M, Rosenshein NB: Mineral fiber exposure and the development of ovarian cancer. Gynecol Oncol 1992;45:20-25.

20. Shushan A, Paltiel O, Iscovich J, Elchalal U, Peretz T, Schenker J: Human menopausal gonadotropin and the risk of epithelial ovarian cancer. Fertil Steril 1996;65:13-18.

21. Tzonou A, Polychronopoulou A., Hsieh CC, Rebelakos A, Karakatsani A, Trichopoulos D: Hair dyes, analgesics, tranquilizers and perineal talc application as risk factors for ovarian cancer. Int J Cancer 1993;55:408-410.

22. Whittemore AS, Wu ML, Paffenbarger RS, Sarles DL, Kampert JB, Grosser S, Jung DL, Ballon S, Hendrickson M: Personal and environmental characteristics related to epithelial ovarian cancer. II. Exposures to talcum powder, tobacco, alcohol, and coffee. Am J Epidemiol 1988;128:1228-1240.

23. Wong C, Hempling RE, Piver MS, Natarajan N, Mettlin CJ: Perineal talc exposure and subsequent epithelial ovarian cancer: A case-control study. Obstet Gynecol 1999;93:372-376.

24. Rothman KJ, Fyler DC, Goldblatt A, Kreidberg MB: Exogenous hormones and other drug exposures of children with congenital heart disease. Am J Epidemiol 1979;109:433-439.

25. Zierler S, Rothman KJ: Congenital heart disease in relation to maternal use of Bendectin and other drugs in early pregnancy. N Engl J Med 1985;313:347-352.

26. Horwitz RI, Feinstein AR: Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med. 1978;299:1089-94.

27. Maclure M: Demonstration of deductive meta-analysis: ethanol intake and risk of myocardial infarction. Epidemiol Rev. 1993;15:328-51.

Appendix

Literature Search Methodology

The literature search was designed to find published epidemiologic studies specifically relating to the perineal use of non-asbestiform talc. The 2000 NTP Draft Report was used as the initial resource to locate applicable studies. To identify other relevant publications, an on-line search was performed in Dialog and using the internet. In addition, medical and scientific resources such as Medline, Toxline, and SciSearch were queried using various keyword terms including "talc," "non-asbestiform," "ovarian cancer," and "perineal." The search was limited to papers published after 1980, because asbestiform products were removed from the market in 1976. Once relevant articles were obtained, bibliographies were "tree-searched" to identify other applicable studies that may have been omitted during the on-line search. "Tree-searching" involves reading an article's bibliography, and then identifying citations that may contain appropriate information based on the title or author. "Tree-searching" identified early studies or those not recorded in on-line databases.

Exhibit 71

Chapter 14

From Association to Causation: Deriving Inferences from Epidemiologic Studies

Not everything that can be counted counts, and not everything that counts can be counted.

—William Bruce Cameron, 1963[1]

Learning Objectives

- To describe a frequent sequence of study designs used to address questions of etiology in human populations.
- To differentiate between real and spurious associations in observational studies.
- To define necessary and sufficient in the context of causal relationships.
- To present guidelines for judging whether an association is causal based on the guidelines set forth by the U.S. Surgeon General, and to discuss the application of these guidelines to broader questions of causal inference.
- To describe how the guidelines for causation originally proposed by the Surgeon General have been modified and utilized by the U.S. Public Health Service and the U.S. Preventive Services Task Force.

In the previous chapters, we discussed the designs of epidemiologic studies that are used to determine whether an association exists between an exposure and a disease (Fig. 14-1A). We then addressed different types of risk measurement that are used to quantitatively express an excess in risk. If we determine that an exposure is associated with a disease, the next question is whether the observed association re! ects a causal relationship (Fig. 14-1B).

Although Figures 14-1A and B refer to an environmental exposure, they could just as well have speci! ed a genetic characteristic or characteristics or a speci! c combination of environmental and genetic factors. As we shall see in Chapter 16, studies of disease etiology generally address the contributions of both genetic and environmental factors and their interactions.

This chapter discusses the derivation of causal inferences in epidemiology. Let us begin by asking, "What approaches are available for studying the etiology of disease?"

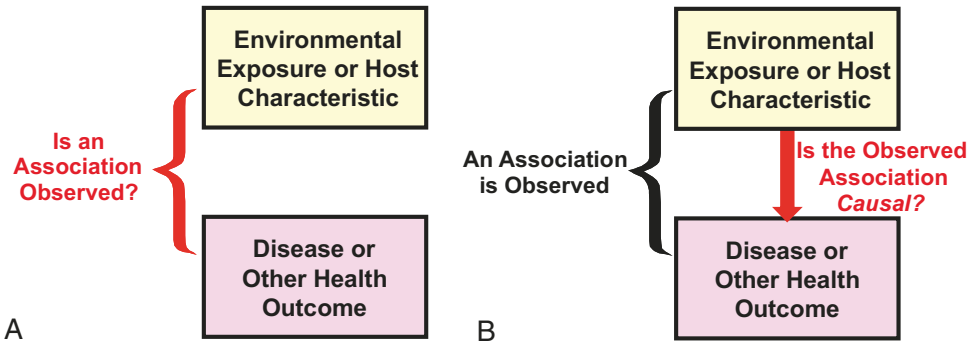


Figure 14-1. **A,** Do we observe an association between exposure and disease? **B,** Is the observed association between exposure and disease causal?

APPROACHES FOR STUDYING DISEASE ETIOLOGY

If we are interested in whether a certain substance is carcinogenic in human beings, a ! rst step in the study of the substance's effect might be to expose animals to the carcinogen in a controlled laboratory environment. Although such animal studies afford us the opportunity to control the exposure dose and other environmental conditions and genetic factors precisely, and to keep loss to follow-up to a minimum, at the conclusion of the study we are left with the problem of having to extrapolate data across species, from animal to human populations. Certain diseases seen in humans have neither occurred nor been produced in animals. It is also dif! cult to extrapolate animal doses to human doses, and species differ in their responses. Thus, although such toxicologic studies can be very useful, they still leave a gnawing uncertainty as to whether the animal ! ndings can be generalized to human beings.

We can also use in vitro systems, such as cell culture or organ culture. However, because these are arti! cial systems, we are again left with the dif- ! culty of extrapolating from arti! cial systems to intact, whole human organisms.

In view of these limitations, if we want to be able to draw a conclusion as to whether a substance causes disease in human beings, we need to make *observations in human populations*. Because we cannot ethically or practically randomize human beings to exposure to a suspected carcinogen, we are dependent on nonrandomized observations, such as those made in case-control and cohort studies.

Approaches to Etiology in Human Populations

Epidemiology capitalizes on what have been called "unplanned" or "natural" experiments. (Some think that this phrase is a contradiction in terms, in that the word "experiment" implies a planned exposure.) What we mean by *unplanned* or *natural* experiments is that we take advantage of groups of people who have been exposed for nonstudy purposes, such as occupational cohorts in speci! c industries or persons exposed to toxic chemicals. Examples include people affected by the poison gas leak disaster at a pesticide manufacturing plant in Bhopal, India, in 1984 and residents of Hiroshima and Nagasaki, Japan, who were exposed to radiation from the atomic bombs dropped on both cities in 1945. Each of these exposed groups can be compared to a nonexposed group to determine whether there is an increased risk of a certain adverse effect in persons who have been exposed.

In conducting human studies, the sequence shown in Figure 14-2 is frequently followed:

The initial step may consist of *clinical observations* at the bedside. For example, when the surgeon Alton Ochsner observed that virtually every patient on whom he operated for lung cancer gave a history of cigarette smoking, he was among the ! rst to suggest a possible causal relationship.[2] A second step is to try to identify *routinely available data*, the analysis of which might shed light on the question. We can then carry out *new studies* such as the cohort and case-control studies discussed in Chapters 9 and 10, which are speci! cally designed to determine whether there is an association between an exposure and a disease, and whether a causal relationship exists.

The usual ! rst step in carrying out new studies to explore a relationship is often a *case-control study*. For example, if Ochsner had wanted to further explore his suggestion that cigarette smoking may be associated with lung cancer, he would have compared the smoking histories of a group of his patients with lung cancer with those of a group of patients without lung cancer—a case-control study.

If a case-control study yields evidence that a certain exposure is suspect, we might next do a *cohort study* (e.g., comparing smokers and nonsmokers and determining the rate of lung cancer in each group or comparing workers exposed to an

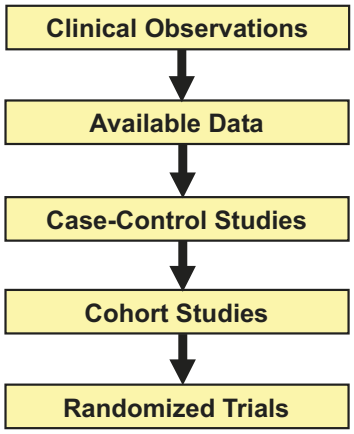


Figure 14-2. A frequent sequence of studies in human populations.

Figure 14-3. Another example of association or causation. (DILBERT © 2011 Scott Adams. Used by permission of UNIVERSAL UCLICK. All rights reserved.)

industrial toxin with workers without such an exposure). Although, in theory, a randomized trial might be the next step, as discussed earlier, randomized trials are almost never used to study the effects of putative toxins or carcinogens and are generally used only for studying potentially beneficial agents.

Conceptually, a two-step process is followed in carrying out studies and evaluating evidence. However, in practice, this process often becomes interactive and deviates from a fixed sequence:

1. We determine whether there is an association or correlation between an exposure or characteristic and the risk of a disease (Fig. 14-3). To do so, we use:
   a. Studies of group characteristics: ecologic studies (discussed in Chapter 10, p. 208)
   b. Studies of individual characteristics: cohort, case-control, and other types of studies
2. If an association is demonstrated, we determine whether the observed association is likely to be a causal one.

TYPES OF ASSOCIATIONS

Real or Spurious Associations

Let us turn next to the types of associations that we might observe in a cohort or case-control study. If we observe an association, we start by asking the question, "Is it a true (real) association or a false (spurious) one?" For example, if we designed a study to select controls in such a way that they tended to be nonexposed, we might observe an association of exposure with disease (i.e., more exposure in cases than in controls). This would not be a true association, but only a result of the study design. Recall that this issue was raised in Chapter 10 regarding a study of coffee consumption and cancer of the pancreas. The possibility was suggested that the controls selected for the study had a lower rate of coffee consumption than was found in the general population.

Interpreting Real Associations

If the observed association is real, is it causal? Figure 14-4 shows two possibilities. Figure 14-4A shows a causal association: we observe an association of exposure and disease, as indicated by the bracket, and the exposure induces development of the disease, as indicated by the arrow. Figure 14-4B shows the same observed association of exposure and disease, but they are associated only because they are both linked to a third factor, designated here as *factor X*. This association is a result of confounding and is noncausal. Confounding is discussed in greater detail in Chapter 15.

In Chapter 10 we discussed this issue in relation to McMahon's study of coffee and cancer of the

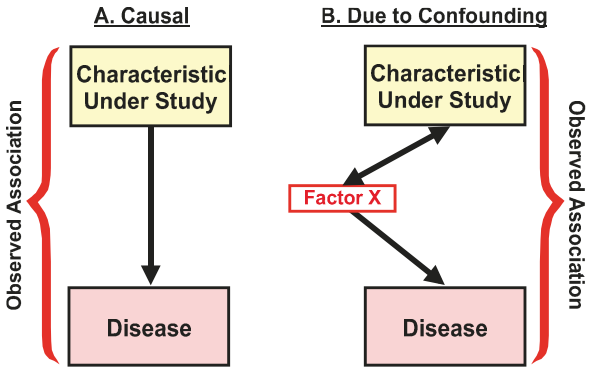


Figure 14-4. Types of associations.

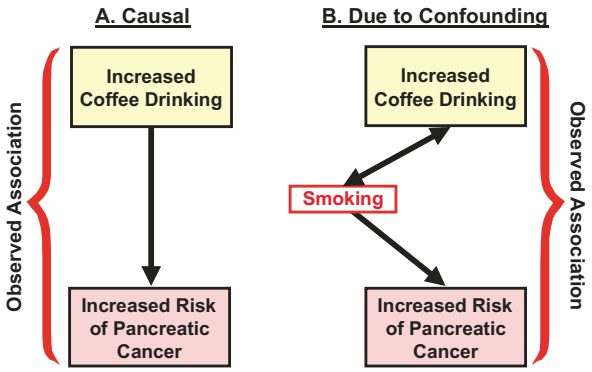


Figure 14-5. Interpreting an observed association between increased coffee drinking and increased risk of pancreatic cancer.

pancreas. McMahon observed an association of coffee consumption with risk of pancreatic cancer. Cigarette smoking was known to be associated with pancreatic cancer, and coffee drinking and cigarette smoking are closely associated (few smokers do not drink coffee) (Fig. 14-5). Therefore, was the observed association of coffee drinking and cancer of the pancreas likely to be a causal relationship, or could the association be due to the fact that coffee and cigarette smoking are associated, and that cigarette smoking is a known risk factor for cancer of the pancreas?

The same issue is exempli! ed by the observed association of increased serum cholesterol level and risk of coronary heart disease (CHD) (Fig. 14-6). Is increased cholesterol a causal factor for increased risk of CHD, or is the observed association due to confounding? That is, are we observing an association of increased cholesterol and CHD because both are associated with a factor X (such as a particular genetic pro! le), which might cause people to have both increased levels of cholesterol and an increased risk of CHD?

Is this distinction really important? What difference does it make? The answer is that it makes a tremendous difference from both clinical and public health standpoints. If the relationship is causal, we will succeed in reducing the risk of CHD if we lower cholesterol levels. However, if the relationship is due to confounding, then the increased risk of CHD is caused by factor X. Therefore, changes in the level of serum cholesterol will have no effect on the risk of CHD. Thus, it is extremely important for us to be able to distinguish between an association due to a causal relationship and an association due to confounding (noncausal).

Let us look at another example. For many years it has been known that cigarette smoking by pregnant women is associated with low birth weight in their infants. As seen in Figure 14-7, the effect is not just the result of the birth of a few low-birth-weight babies in this group of women. Rather, the entire weight distribution curve is shifted to the left in the babies born to smokers. The reduction in birth weight is also not a result of shorter pregnancies. The babies of smokers are smaller than those of nonsmokers at each gestational age (Fig. 14-8). A dose-response relationship is also seen (Fig. 14-9). The more a woman smokes, the greater her risk of having a low-birth-weight baby. For many years the interpretation of this association was the subject of great controversy. Many believed the association re! ected a causal relation. Others, including a leading statistician, Jacob Yerushalmy, believed the association was due to confounding and was not causal. He wrote as follows:

> *A comparison of smokers and nonsmokers shows that the two differ markedly along many environmental, behavioral and biologic variables. For example, smokers are less likely to use contraceptives and to plan the pregnancy. Smokers are more likely to drink coffee, beer and whiskey and the nonsmoker, tea, milk and wine. The smoker is more likely than the nonsmoker to indulge in these habits to excess. In general, the nonsmokers are revealed to be more moderate than the smokers who are shown to be more extreme and carefree in their mode of life. Some biologic differences are also noted between them: Thus smokers have a higher twinning rate only in whites and their age for menarche is lower than for nonsmokers.*[3]

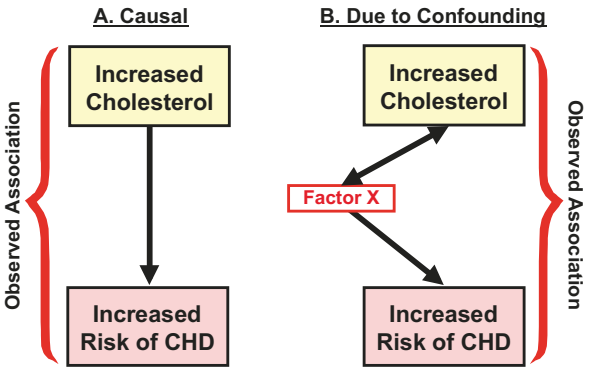


Figure 14-6. Interpreting an observed association between increased cholesterol level and increased risk of coronary heart disease (CHD).

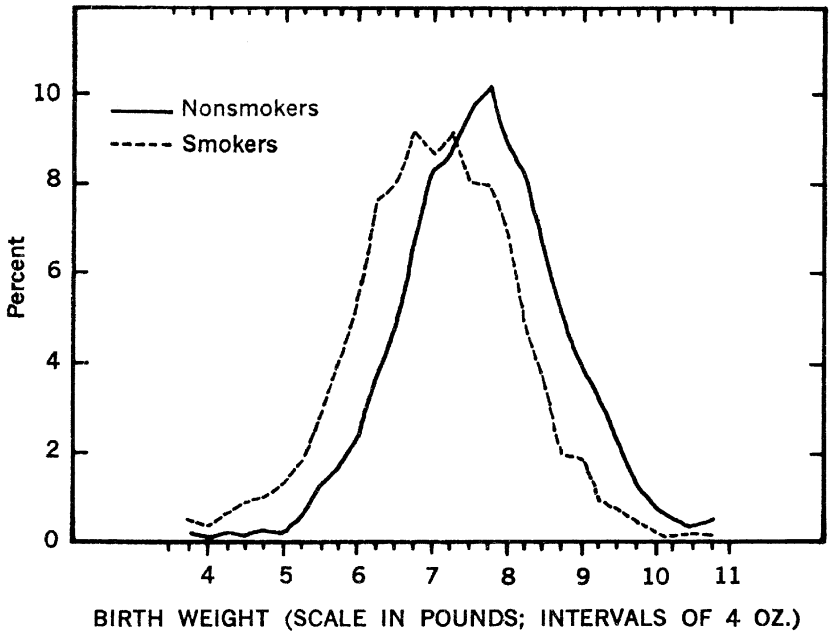


Figure 14-7. Percentage distribution by birth weight of infants of mothers who did not smoke during pregnancy and of those mothers who smoked 1 pack of cigarettes or more per day. (From U.S. Department of Health, Education, and Welfare: The Health Consequences of Smoking. Washington, DC, Public Health Service, 1973, p 105.)

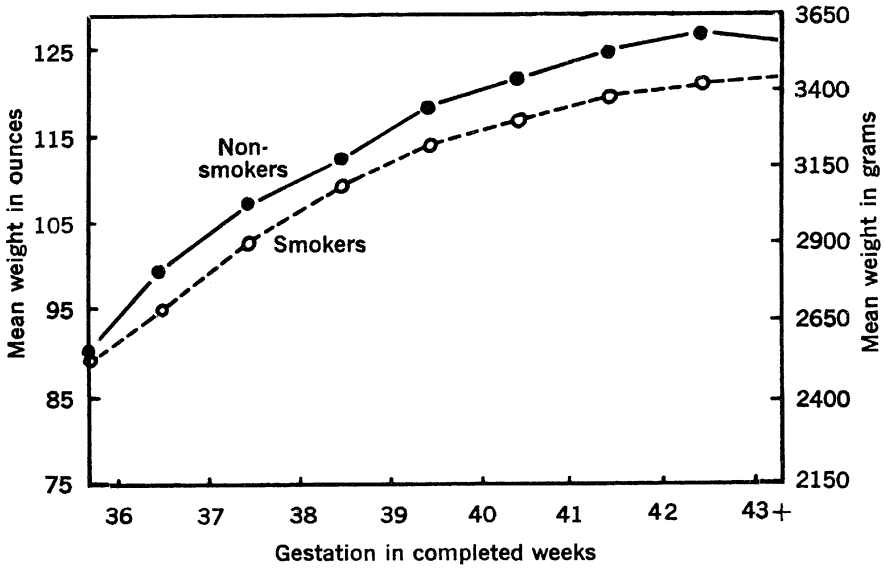


Figure 14-8. Mean birth weight for week of gestation according to maternal smoking habit. (From U.S. Department of Health, Education, and Welfare: The Health Consequences of Smoking. Washington, DC, Public Health Service, 1973, p 104.)

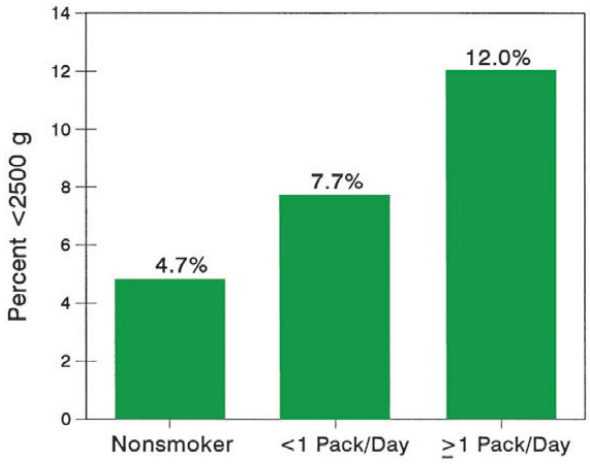


Figure 14-9. Percentage of pregnancies (n = 50,267) with infant weighing less than 2,500 g, by maternal cigarette smoking category. (Redrawn from Ontario Department of Health: Second Report of the Perinatal Mortality Study in Ten University Teaching Hospitals. Toronto, Ontario, Department of Health, Ontario Perinatal Mortality Study Committee, Vol. I, 1967, p 275.)

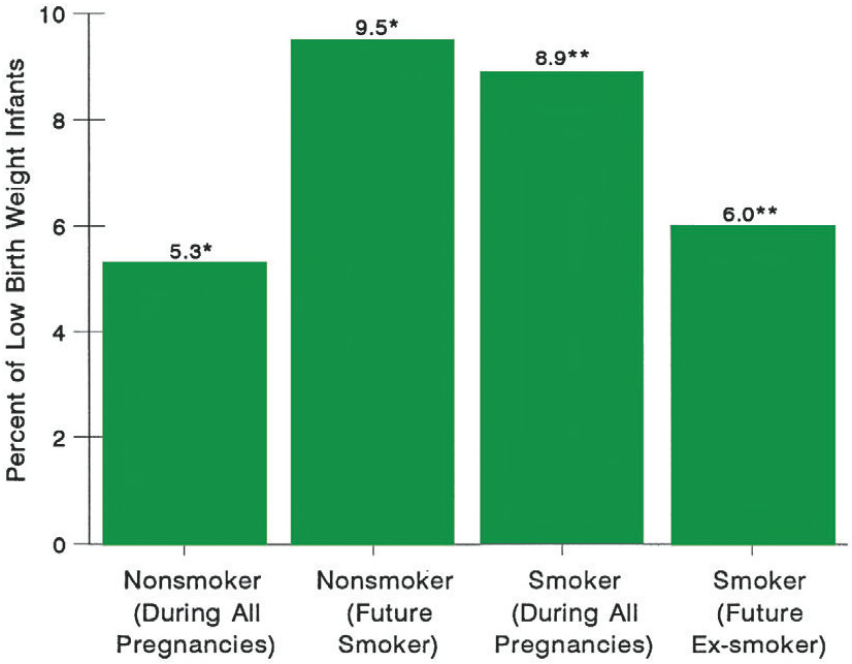


Figure 14-10. Percentage of low-birth-weight infants by smoking status of their mothers (*$P < .01$; **$P < .02$). (Redrawn from Yerushalmy J: Infants with low birth weight born before their mothers started to smoke cigarettes. Am J Obstet Gynecol 112:277–284, 1972.)

In view of these many differences between smokers and nonsmokers, Yerushalmy believed that it was not the *smoking* that caused the low birth weight, but rather that the low weight was attributable to *other characteristics of the smokers*. It is interesting to examine a study that Yerushalmy carried out to support his position at the time (Fig. 14-10).[3]

Yerushalmy examined the results of one pregnancy (the study pregnancy) in a population of women who had had several pregnancies. The rate of low-birth-weight babies in the study pregnancy was 5.3% for women who were nonsmokers in *all* of their pregnancies. However, if they were smokers in all of their pregnancies, the rate of low birth weight in the study pregnancy was almost 9%. When he examined pregnancies of women who were nonsmokers during the study pregnancy, but who later became smokers, he found that their rate of low-birth-weight babies was about equal to that of women who smoked in all pregnancies. When he examined pregnancies of women who were smokers in the study pregnancy, but who subsequently stopped smoking, he found that their rate of low birth weight in the study pregnancy was similar to that of women who were nonsmokers in all of their pregnancies.

On the basis of these data, Yerushalmy came to the conclusion that it was not the smoking but rather some characteristic of the smoker that caused the low birth weight. Today, however, it is virtually universally accepted that smoking is a cause of low birth weight. The causal nature of this relation has also been demonstrated in randomized trials that have reduced the frequency of low birth weight by initiating programs for smoking cessation in pregnant women. Although this issue has now largely been resolved, it is illuminating to review both the controversy and the study, as they exemplify the reasoning that is necessary in trying to distinguish causal from noncausal interpretations of observed associations.

TYPES OF CAUSAL RELATIONSHIPS

A causal pathway can be either *direct* or *indirect* (Fig. 14-11). In *direct* causation, a factor directly

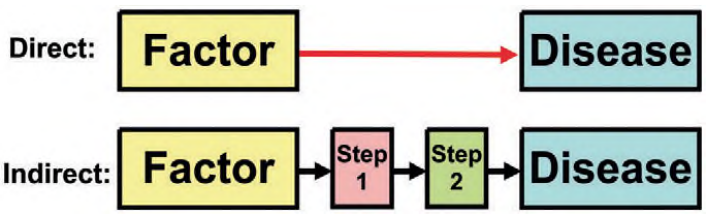


Figure 14-11. Direct versus indirect causes of disease.

Figure 14-12. Types of causal relationships: I. Factor A is both necessary and suf! cient.

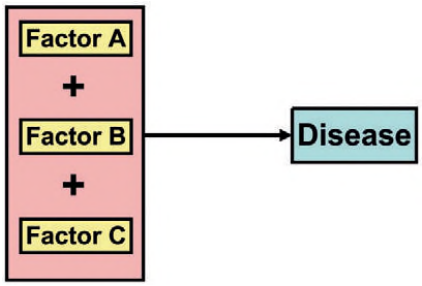


Figure 14-13. Types of causal relationships: II. Each factor is necessary, but not suf! cient.

causes a disease without any intermediate step. In *indirect* causation, a factor causes a disease, but only through an intermediate step or steps. In human biology, intermediate steps are virtually always present in any causal process.

If a relationship is causal, four types of causal relationships are possible: (1) necessary and suf! cient; (2) necessary, but not suf! cient; (3) suf! cient, but not necessary; and (4) neither suf! cient nor necessary.

Necessary and Sufficient

In the ! rst type of causal relationship, a factor is both necessary and suf! cient for producing the disease. Without that factor, the disease never develops (the factor is necessary), and in the presence of that factor, the disease always develops (the factor is suf! cient) (Fig. 14-12). This situation rarely if ever occurs. For example, in most infectious diseases, a number of people are exposed, some of whom will manifest the disease and others who will not. Members of households of a person with tuberculosis do not uniformly acquire the disease from the index case. If the exposure dose is assumed to be the same, there are likely differences in immune status, genetic susceptibility, or other characteristics that determine who develops the disease and who does not. A one-to-one relationship of exposure to disease, which is a consequence of a necessary and suf! cient relationship, rarely if ever occurs.

Necessary, But Not Sufficient

In another model, each factor is necessary, but not, in itself, suf! cient to cause the disease (Fig. 14-13). Thus, multiple factors are required, often in a speci! c temporal sequence. For example,

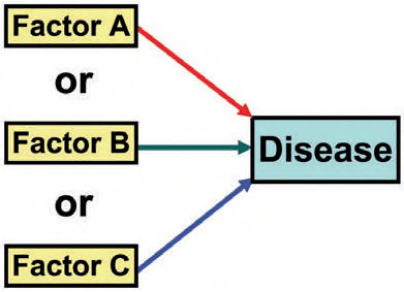


Figure 14-14. Types of causal relationships: III. Each factor is suf! cient, but not necessary.

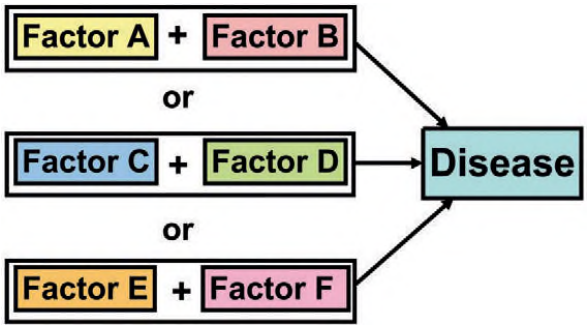


Figure 14-15. Types of causal relationships: IV. Each factor is neither suf! cient nor necessary.

carcinogenesis is considered to be a multistage process involving both initiation and promotion. For cancer to result, a promoter must act after an initiator has acted. Action of an initiator or a promoter alone will not produce a cancer.

Again, in tuberculosis, the tubercle bacillus is clearly a necessary factor, even though its presence may not be suf! cient to produce the disease in every infected individual.

Sufficient, But Not Necessary

In this model, the factor alone can produce the disease, but so can other factors that are acting alone (Fig. 14-14). Thus, either radiation exposure or benzene exposure can each produce leukemia without the presence of the other. Even in this situation, however, cancer does not develop in everyone who has experienced radiation or benzene exposure, so although both factors are not needed, other cofactors probably are. Thus, the criterion of *suf! cient* is rarely met by a single factor.

Neither Sufficient Nor Necessary

In the fourth model, a factor, by itself, is neither suf! cient nor necessary to produce disease (Fig. 14-15). This is a more complex model, which probably most accurately represents the causal relationships that operate in most chronic diseases.

EVIDENCE FOR A CAUSAL RELATIONSHIP

Many years ago, when the major disease problems faced by man were infectious in origin, the question arose as to what evidence would be necessary to prove that an organism causes a disease. In 1840, Henle proposed postulates for causation that were expanded by Koch in the 1880s.[4] The postulates for causation were as follows:

1. The organism is *always* found with the disease.
2. The organism is *not* found with any other disease.
3. The organism, when isolated from one who has the disease, and cultured through several generations, produces the disease (in experimental animals).

Koch added that "Even when an infectious disease cannot be transmitted to animals, the 'regular' and 'exclusive' presence of the organism [postulates 1 and 2] proves a causal relationship."[4]

These postulates, though not perfect, proved very useful for infectious diseases. However, as apparently noninfectious diseases assumed increasing importance toward the middle of the 20th century, the issue arose as to what would represent strong evidence of causation in diseases that were generally not of infectious origin. In such disease there was no organism that could be cultured and grown in animals. Speci! cally, as attention was directed to a possible relationship between smoking and lung cancer, the U.S. Surgeon General appointed an expert committee to review the evidence. The committee developed a set of guidelines,[5] which have been revised over the years. The next few pages present a modi! ed list of these guidelines (Table 14-1) with some brief comments.

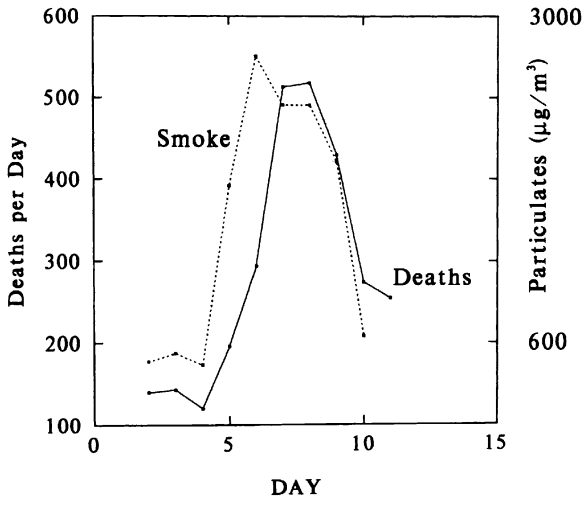


Figure 14-16. The mean concentration of airborne particles (µg/m³) from the four inner monitoring stations in London and the count of daily deaths in the London Administrative County during the beginning of December 1952. (From Schwartz J: Air pollution and daily mortality: A review and meta analysis. Environ Res 64:36–52, 1994.)

TABLE 14-1. **Guidelines for Judging Whether an Observed Association Is Causal**

1. Temporal relationship
2. Strength of the association
3. Dose-response relationship
4. Replication of the ! ndings
5. Biologic plausibility
6. Consideration of alternate explanations
7. Cessation of exposure
8. Consistency with other knowledge
9. Speci! city of the association

GUIDELINES FOR JUDGING WHETHER AN OBSERVED ASSOCIATION IS CAUSAL

1. Temporal Relationship. It is clear that if a factor is believed to be the cause of a disease, exposure to the factor must have occurred before the disease developed. Figure 14-16 shows the number of deaths per day and the mean concentration of airborne particles in London in early December 1952.[6] The pattern of a rise in particle concentration followed by a rise in mortality and a subsequent decline in particle concentration followed by a decline in mortality strongly supported the increase in mortality being due to the increase in air pollution. This example demonstrates the use of ecologic data for exploring a temporal relationship. Further investigation revealed that the increased mortality consisted almost entirely of respiratory and cardiovascular deaths and was highest in the elderly.

It is often easier to establish a temporal relationship in a prospective cohort study than in a case-control study or a retrospective cohort study. In the last two types of studies, exposure information may need to be obtained or re-created from past records and the timing may therefore be imprecise.

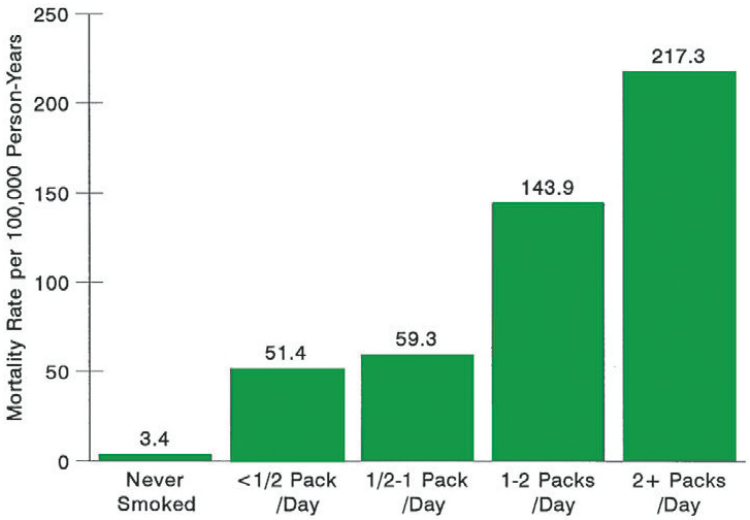


Figure 14-17. Age-standardized death rates due to well-established cases of bronchogenic carcinoma (exclusive of adenocarcinoma) by current amount of smoking. (Adapted from Hammond EC, Horn D: Smoking and death rates: Report on 44 months of follow-up of 187,783 men: II. Death rates by cause. JAMA 166:1294–1508, 1958. Copyright 1958, American Medical Association.)

The temporal relationship of exposure and disease is important not only for clarifying the order in which the two occur but also in regard to the length of the interval between exposure and disease. For example, asbestos has been clearly linked to increased risk of lung cancer, but the latent period between the exposure and the appearance of lung cancer is at least 15 to 20 years. Therefore, if, for example, lung cancer develops after only 3 years since the asbestos exposure, it is safe to conclude that the lung cancer was not a result of this exposure.

2. Strength of the Association. The strength of the association is measured by the relative risk (or odds ratio). The stronger the association, the more likely it is that the relation is causal.

3. Dose-Response Relationship. As the dose of exposure increases, the risk of disease also increases. Figure 14-17 shows an example of the dose-response relationship for cigarette smoking and lung cancer. If a dose-response relationship is present, it is strong evidence for a causal relationship. However, the absence of a dose-response relationship does not necessarily rule out a causal relationship. In some cases in which a threshold may exist, no disease may develop up to a certain level of exposure (a threshold); above this level, disease may develop.

4. Replication of the Findings. If the relationship is causal, we would expect to find it consistently in different studies and in different populations. Replication of findings is particularly important in epidemiology. If an association is observed, we would also expect it to be seen consistently within subgroups of the population and in different populations, unless there is a clear reason to expect different results.

5. Biologic Plausibility. Biologic plausibility refers to coherence with the current body of biologic knowledge. Examples may be cited to demonstrate that epidemiologic observations have sometimes preceded biologic knowledge. Thus, as discussed in an earlier chapter, Gregg's observations on rubella and congenital cataracts preceded any knowledge of teratogenic viruses. Similarly, the implication of high oxygen concentration in the causation of retrolental fibroplasia, a form of blindness that occurs in premature infants, preceded any biologic knowledge supporting such a relationship. Nevertheless, we seek consistency of the epidemiologic findings with existing biologic knowledge, and when this is not the case, interpreting the meaning of the observed association may be difficult. We may then be more demanding in our requirements about the size and significance of any differences observed and in having the study replicated by other investigators in other populations.

6. Consideration of Alternate Explanations. We have discussed the problem in interpreting an observed association in regard to whether a relationship is causal or is the result of confounding. In judging whether a reported association is causal, the extent to which the investigators have taken other possible explanations into account and the extent to which they have ruled out such explanations are important considerations.

7. Cessation of Exposure. If a factor is a cause of a disease, we would expect the risk of the disease to decline when exposure to the factor is reduced

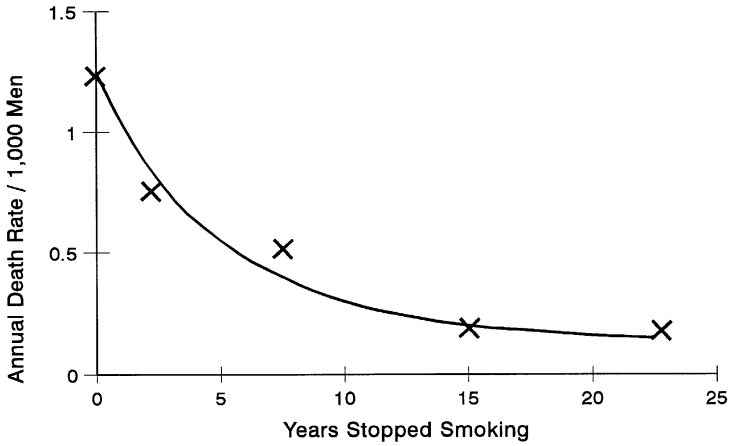


Figure 14-18. Effects of terminating exposure: lung cancer death rates, standardized for age and amount smoked, among men continuing to smoke cigarettes and men who gave up smoking for different periods. The corresponding rate for nonsmokers was 0.07 per 1,000. (Adapted from Doll R, Hill AB: Mortality in relation to smoking: Ten years' observations of British doctors. BMJ 1:1399–1410, 1964.)

or eliminated. Figure 14-18 shows such data for cigarette smoking and lung cancer.

Eosinophilia-myalgia syndrome (EMS) reached epidemic proportions in 1989. Characterized by severe muscle pain and a high blood eosinophil count, the syndrome was found to be associated with manufactured preparations of L-tryptophan. In November 1989, a nationwide recall by the Food and Drug Administration of over-the-counter preparations of L-tryptophan was followed by dramatic reductions in numbers of cases of EMS reported each month (Fig. 14-19). This is another example of a reduction in incidence being related to cessation of exposure, which adds to the strength of the causal inference regarding the exposure.

When cessation data are available, they provide helpful supporting evidence for a causal association. However, in certain cases, the pathogenic process may have been irreversibly initiated, and the disease occurrence may have been determined by the time the exposure is removed. Emphysema is not reversed with cessation of smoking, but its progression is reduced.

8. Consistency with Other Knowledge. If a relationship is causal, we would expect the ! ndings to be consistent with other data. For example, Figure 14-20 shows data regarding lung cancer rates in men and women and cigarette smoking in men and women.

We see a consistent direction in the curves, with the increase in lung cancer rates following the increase in cigarette sales in both men and women. These data are consistent with what we would expect if the relationship between smoking and lung cancer is established as a causal one. Although the absence of such consistency would not completely rule out this hypothesis, if we observed rising lung cancer rates after a period of declining

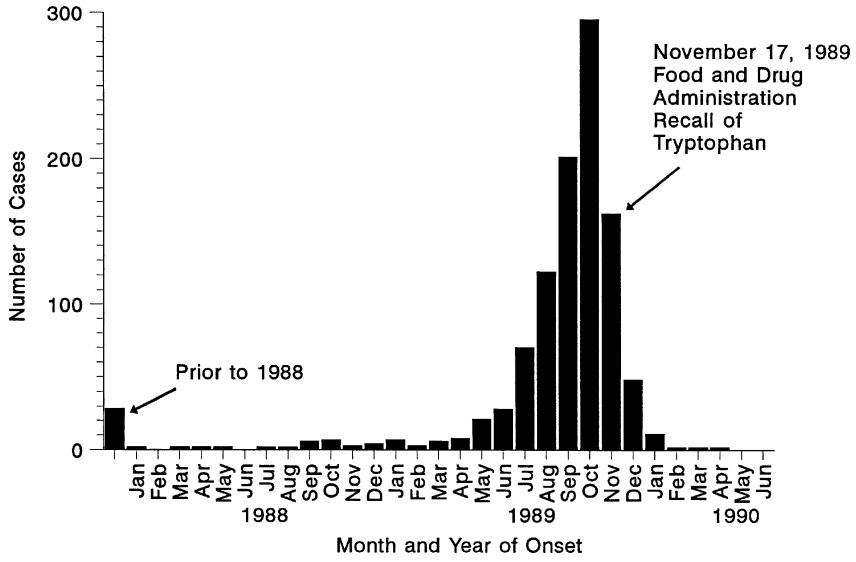


Figure 14-19. Reported dates of illness onset by month and year for cases of eosinophilia-myalgia syndrome, as reported to the Centers for Disease Control and Prevention, Atlanta, as of July 10, 1990. (Adapted from Swygert LA, Maes EF, Sewell LE, et al: Eosinophilia-myalgia syndrome: Results of national surveillance. JAMA 264:1698–1703, 1990. Copyright 1990, American Medical Association.)

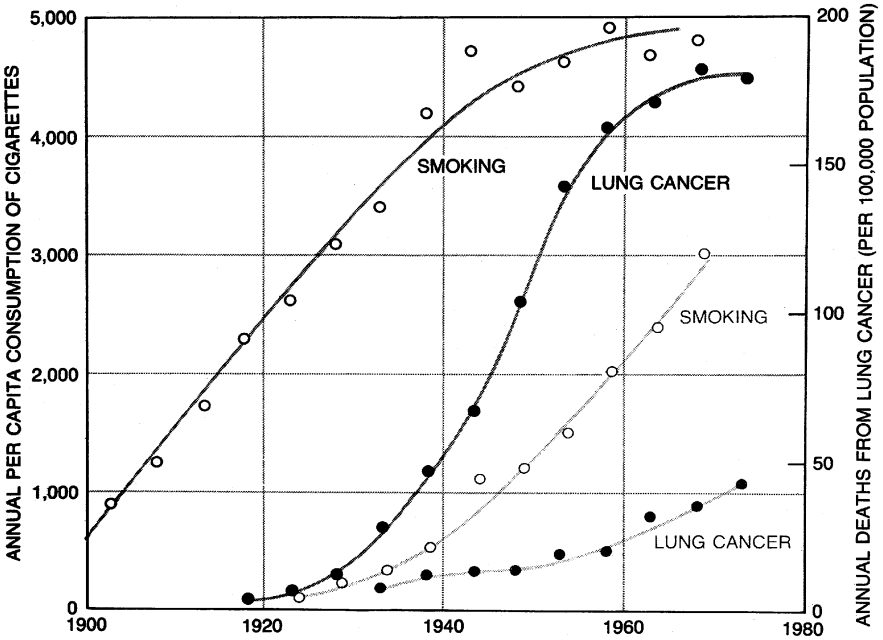


Figure 14-20. Parallel trends between cigarette consumption and lung cancer in men (two curves on left) and in women (two curves on right), in England and Wales. (From Cairns J: The cancer problem. Sci Am 233:64–72, 77–78, 1975.)

cigarette sales, for example, we would need to explain how this observation could be consistent with a causal hypothesis.

9. Specificity of the Association. An association is speci! c when a certain exposure is associated with only one disease; this is the weakest of all the guidelines and should probably be deleted from the list. Cigarette manufacturers have pointed out that the diseases attributed to cigarette smoking do not meet the requirements of this guideline, because cigarette smoking has been linked to lung cancer, pancreatic cancer, bladder cancer, heart disease, emphysema, and other conditions.

The possibility of such multiple effects from a single factor is not, in fact, surprising: regardless of the tissue that comprises them, all cells have common characteristics, including DNA, RNA, and various subcellular structures, so a single agent could have effects in multiple tissues. Furthermore, cigarettes are not a single factor but constitute a mixture of a large number of compounds; consequently, a large number of effects might be anticipated.

When speci! city of an association is found, it provides additional support for a causal inference. However, as with a dose-response relationship, absence of speci! city in no way negates a causal relationship.

Any conclusion that an observed association is causal is greatly strengthened when different types of evidence from multiple sources support such reasoning. Thus, it is not so much a count of the number of guidelines present that is relevant to causal inference but rather an *assessment of the total pattern of evidence observed* that may be consistent with one or more of the guidelines. Sir Austin Bradford Hill eloquently expressed this sentiment in an essay written in 1965:

> *Here then are nine different viewpoints [guidelines] from all of which we should study association before we cry causation. What I do not believe—and this has been suggested—that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?*[7]

DERIVING CAUSAL INFERENCES: TWO EXAMPLES

Peptic Ulcers and Gastric Cancer in Relation to Infection with *Helicobacter pylori*

Although the preceding guidelines do not permit a quantitative estimation of whether or not an

association is causal, they can nevertheless be very helpful, as seen in the following examples:

Until the 1980s, the major causes of peptic ulcer disease were considered to be stress and lifestyle factors, including smoking. Peptic ulcer disease had long been attributed to the effects of gastric acid. Susceptibility to gastric acid had been linked to cigarette smoking, alcohol consumption, and use of nonsteroidal anti-in! ammatory agents. Therapy was primarily directed at inhibiting acid secretion and protecting mucosal surfaces from acid. Although these therapies helped healing, relapses were common.

In 1984, Australian physicians Drs. Barry J. Marshall and J. Robin Warren reported that they had observed small curved bacteria colonizing the lower part of the stomach in patients with gastritis and peptic ulcers.[8] After several attempts, Marshall succeeded in cultivating a hitherto unknown bacterial species (later named *Helicobacter pylori*) from several of these biopsies (Fig. 14-21). Together they found that the organism was present in almost all patients with gastric in! ammation or peptic ulcer. Many of these patients had biopsies performed which showed evidence of in! ammation present in the gastric mucosa close to where the bacteria were seen. Based on these results, they proposed that *Helicobacter pylori* is involved in the etiology of these diseases. It was subsequently shown that the ulcer was often not cured until *Helicobacter pylori* had been eliminated.

It is now ! rmly established that *Helicobacter pylori* causes more than 90% of duodenal ulcers and up to 80% of gastric ulcers. The link between *Helicobacter pylori* infection and subsequent gastritis and peptic ulcer disease has been established through studies of human volunteers, antibiotic treatment studies, and epidemiological studies. Thus, many of the study designs discussed in previous chapters and many of the guidelines for causal inferences discussed earlier in this chapter were involved in elucidating the role of *Helicobacter pylori* in peptic ulcer and gastritis. In 2005, the Nobel Prize for Physiology or Medicine was shared by Drs. Marshall and Warren, "for their discovery of the bacterium *Helicobacter pylori* and its role in gastritis and peptic ulcer disease."

Table 14-2 categorizes this evidence according to several of the guidelines for causation just discussed. Thus, as seen here, the guidelines can be extremely helpful in characterizing the evidence supporting a causal relationship.

Figure 14-21. *Helicobacter pylori* [Photograph]. (Encyclopædia Britannica Online. http://www.britannica.com/EBchecked/topic/450889/peptic-ulcer?overlay=true&assemblyId=94921. Accessed August 15, 2013.)

Increasing evidence now also supports the association of *Helicobacter pylori* infection and the development of gastric cancer. Uemura and coworkers[9] prospectively studied 1,526 Japanese patients who had duodenal or gastric ulcers, gastric hyperplasia, or nonulcer hyperplasia. Of this group, 1,246 had *Helicobacter pylori* infection and 280 did not. The mean follow-up period was 7.8 years. Gastric cancers developed in 36 (2.9%) of the infected patients, but in none of the noninfected patients. Individuals who carry antibodies to *Helicobacter pylori* may have a 2 to 3 times higher risk of stomach cancer than those who do not (Fig. 14-22). The risk of stomach cancer also appears to be related to the type of strain of *Helicobacter pylori* which is infecting a person. Evidence is accumulating to support the idea that therapy against *Helicobacter pylori* may prevent gastric cancer. In the future, gastric cancer may come to be viewed as a largely preventable cancer of infectious origin.

Age of Onset of Alcohol Use and Lifetime Alcohol Abuse

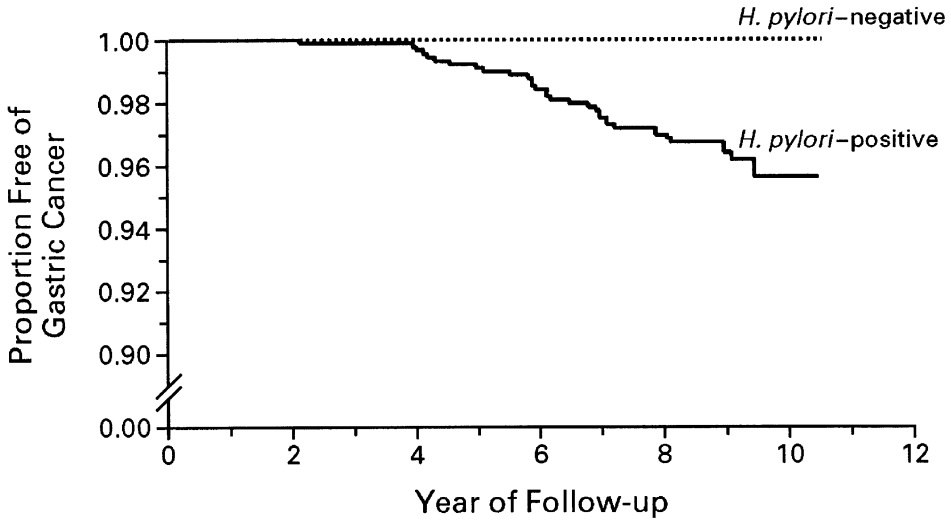
In 1997, Grant and Dawson[10] reported data on the relationship of age at ! rst use of alcohol and prevalence of lifetime alcohol dependence and abuse.

TABLE 14-2. **Assessment of the Evidence Suggesting *Helicobacter pylori* as a Causative Agent of Duodenal Ulcers**

1. **Temporal relationship.**
   - *Helicobacter pylori* is clearly linked to chronic gastritis. About 11% of chronic gastritis patients will go on to have duodenal ulcers over a 10-year period.
   - In one study of 454 patients who underwent endoscopy 10 years earlier, 34 of 321 patients who had been positive for *Helicobacter pylori* (11%) had duodenal ulcer compared with 1 of 133 *Helicobacter pylori*–negative patients (0.8%).
2. **Strength of the association.**
   - *Helicobacter pylori* is found in at least 90% of patients with duodenal ulcer. In at least one population reported to lack duodenal ulcers, a northern Australian aboriginal tribe that is isolated from other people, it has never been found.
3. **Dose-response relationship.**
   - Density of *Helicobacter pylori* per square millimeter of gastric mucosa is higher in patients with duodenal ulcer than in patients without duodenal ulcer. Also see item 2 above.
4. **Replication of the findings.**
   - Many of the observations regarding *Helicobacter pylori* have been replicated repeatedly.
5. **Biologic plausibility.**
   - Although originally it was dif! cult to envision a bacterium that infects the stomach antrum causing ulcers in the duodenum, it is now recognized that *Helicobacter pylori* has binding sites on antral cells and can follow these cells into the duodenum.
   - *Helicobacter pylori* also induces mediators of in! ammation.
   - *Helicobacter pylori*–infected mucosa is weakened and is susceptible to the damaging effects of acid.
6. **Consideration of alternate explanations.**
   - Data suggest that smoking can increase the risk of duodenal ulcer in *Helicobacter pylori*–infected patients but is not a risk factor in patients in whom *Helicobacter pylori* has been eradicated.
7. **Cessation of exposure.**
   - Eradication of *Helicobacter pylori* heals duodenal ulcers at the same rate as histamine receptor antagonists.
   - Long-term ulcer recurrence rates were zero after *Helicobacter pylori* was eradicated using triple-antimicrobial therapy, compared with a 60% to 80% relapse rate often found in patients with duodenal ulcers treated with histamine receptor antagonists.
8. **Consistency with other knowledge.**
   - Prevalence of *Helicobacter pylori* infection is the same in men as in women. The incidence of duodenal ulcer, which in earlier years was believed to be higher in men than in women, has been equal in recent years.
   - The prevalence of ulcer disease is believed to have peaked in the latter part of the 19th century, and the prevalence of *Helicobacter pylori* may have been much higher at that time because of poor living conditions. This reasoning is also based on observations today that the prevalence of *Helicobacter pylori* is much higher in developing countries.
9. **Specificity of the association.**
   - Prevalence of *Helicobacter pylori* in patients with duodenal ulcers is 90% to 100%. However, it is found in some patients with gastric ulcer and even in asymptomatic individuals.

Data from Megraud F, Lamouliatte H: *Helicobacter pylori* and duodenal ulcer: Evidence suggesting causation. Dig Dis Sci 37:769–772, 1992; and DeCross AJ, Marshall BJ: The role of *Helicobacter pylori* in acid-peptic disease. Am J Med Sci 306: 381–392, 1993.

They analyzed data from 27,616 current and former drinkers who were interviewed as part of the 1992 National Longitudinal Alcohol Epidemiologic Survey. The rates of lifetime *dependence* decreased from more than 40% among individuals who began drinking at age 14 years or younger to about 10% among those who started drinking at age 20 years or older (Fig. 14-23). The con! guration of the curve in Figure 14-23 suggests a dose-response relationship as has been observed for longer duration of smoking associated with increased risk of lung cancer. However, the data may also point to a period of particularly high susceptibility, namely, that the period of preadolescence and early adolescence is a period of increased risk for developing a disorder of alcohol use. Therefore, interventions



No. at Risk						
H. pylori–negative	280	272	251	245	213	57
H. pylori–positive	1246	1219	1086	907	782	258

Figure 14-22. Kaplan-Meier analysis of the proportion of *Helicobacter pylori*–positive and *Helicobacter pylori*–negative patients who remained free of gastric cancer. During follow-up, gastric cancer developed in 36 of the 1,246 *H. pylori*–infected patients (2.9%), but in none of the 280 uninfected patients ($P < .001$). (From Uemura N, Okamoto S, Yamomoto S, et al: *Helicobacter pylori* infection and the development of gastric cancer. N Engl J Med 345:784–789, 2001.)

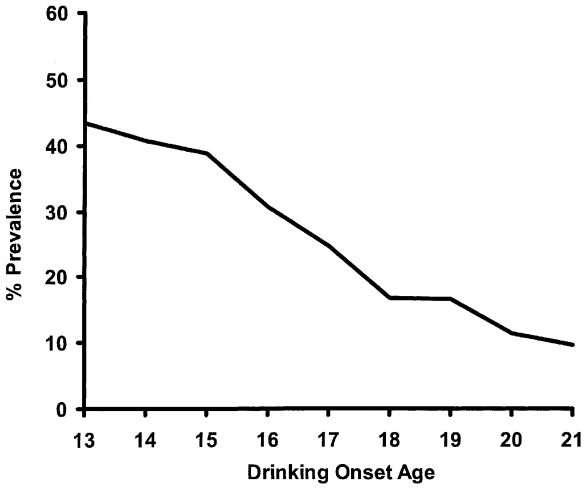


Figure 14-23. Relation of age of onset of alcohol use to prevalence of lifetime alcohol abuse. (Adapted from Grant BF, Dawson DA: Age at onset of alcohol use and its association with DSM-IV alcohol abuse and dependence: Results from the National Longitudinal Alcohol Epidemiologic Survey. J Subst Abuse 9:103–110, 1997.)

should be targeted to this group in the hope of delaying drinking onset. However, adopting such an approach assumes that the relationship between early onset of drinking and subsequent lifetime abuse is a causal one, so that delaying age at onset of drinking would reduce the risk of lifetime alcohol dependence. Another possible explanation is that those who are destined for lifetime alcohol dependence tend to begin drinking earlier, but that the earlier age at drinking onset is not necessarily a cause of the later dependence. Further research is therefore needed to explain the intriguing association that has been observed. We shall return to this example in Chapter 16.

MODIFICATIONS OF THE GUIDELINES FOR CAUSAL INFERENCES

In 1986, the U.S. Public Health Service brought together a group of 19 experts to examine the scienti! c basis of the content of prenatal care and to answer the question: Which measures implemented during prenatal care have actually been demonstrated to be associated with improved outcome? The panel's report was issued in 1989 and served as the basis of a comprehensive report.[11] As the panel began its deliberations, it became clear that questions of causation were at the heart of the panel's task, and that guidelines were needed for assessing the relationship of prenatal measures to health outcomes. A subcommittee reviewed the current guidelines (just enumerated in the preceding text) and de! ned a process for using evidence that includes (1) categorization of the evidence by the

TABLE 14-3. **The Process for Using the Evidence in Developing Recommendations on the Effectiveness of Prenatal Interventions**

Stage I: Categorizing the Evidence by the Quality of Its Source. (*In each category, studies are listed in descending order of quality.*)

1. Trials (planned interventions with contemporaneous assignment of treatment and nontreatment)
   a. Randomized, double-blind, placebo-controlled with sufficient power appropriately analyzed.
   b. Randomized, but blindness not achieved.
   c. Nonrandomized trials with good control of confounding, that are well conducted in other respects.
   d. Randomized, but with deficiencies in execution or analysis (insufficient power, major losses to follow-up, suspect randomization, analysis with exclusions).
   e. Nonrandomized trials with deficiencies in execution or analysis.
2. Cohort or case-control studies
   a. Hypothesis specified before analysis, good data, confounders accounted for.
   b. As above, but hypothesis not specified before analysis.
   c. Post hoc, with problem(s) in the data or the analysis.
3. Time-series studies
   a. Analyses that take confounding into account.
   b. Analyses that do not consider confounding.
4. Case-series studies: Series of case reports without any specific comparison group

Among other issues that must be considered in reviewing the evidence are the precision of definition of the outcome being measured, the degree to which the study methodology has been described, adequacy of the sample size, and the degree to which characteristics of the population studied and of the intervention being evaluated have been described.

A study can be well designed and carried out in an exemplary fashion (internal validity), but if the population studied is an unusual or highly selected one, the results may not be generalizable (external validity).

Stage II: Guidelines for Evaluating the Evidence of a Causal Relationship. (*In each category, studies are listed in descending priority order.*)

1. Major criteria
   a. Temporal relationship: An intervention can be considered evidence of a reduction in risk of disease or abnormality only if the intervention was applied before the time the disease or abnormality would have developed.
   b. Biological plausibility: A biologically plausible mechanism should be able to explain why such a relationship would be expected to occur.
   c. Consistency: Single studies are rarely definitive. Study findings that are replicated in different populations and by different investigators carry more weight than those that are not. If the findings of studies are inconsistent, the inconsistency must be explained.
   d. Alternative explanations (confounding): The extent to which alternative explanations have been explored is an important criterion in judging causality.
2. Other considerations
   a. Dose-response relationship: If a factor is indeed the cause of a disease, usually (but not invariably) the greater the exposure to the factor, the greater the risk of the disease. Such a dose-response relationship may not always be seen because many important biologic relationships are dichotomous, and reach a threshold level for observed effects.
   b. Strength of the association: The strength of the association is usually measured by the extent to which the relative risk or odds depart from unity, either above 1 (in the case of disease-causing exposures) or below 1 (in the case of preventive interventions).
   c. Cessation effects: If an intervention has a beneficial effect, then the benefit should cease when it is removed from a population (unless carryover effect is operant).

Adapted from Gordis L, Kleinman JC, Klerman LV, et al: Criteria for evaluating evidence regarding the effectiveness of prenatal interventions. In Merkatz IR, Thompson JE (eds): New Perspectives on Prenatal Care. New York, Elsevier, 1990, pp 31–38.

quality of its sources, and (2) evaluation of the evidence of a causal relationship using standardized guidelines.[12] These recommendations are excerpted in Table 14-3. Although these modi! ed guidelines clearly use the original components, they establish reasonable priorities in weighting them. They thus de! ne an approach for looking at causation that may have applicability far beyond questions of the effectiveness of prenatal measures.

A similar approach, ranking studies by the quality of the study and its evidence, is used by the U.S. Preventive Services Task Force, which is responsible for developing clinical practice guidelines for prevention and screening (Table 14-4).[13] The Task Force is an independent committee of experts supported by the U.S. Government. Members include experts in primary care, prevention, evidence-based medicine, and research methods. Various clinical areas and experience in preventive medicine, public health, and health policy are also represented.

For each topic the Task Force considers, it de! nes the questions that need to be addressed and identi! es and retrieves the relevant evidence. The quality of each individual study is assessed after which the strength of the totality of available evidence is judged. Estimates are made of the balance of bene! ts and harms. This balance is expressed as the net bene! t (the difference between bene! ts and harms). The Task Force prepares recommendations for preventive interventions based on these considerations.

Figure 14-24 shows a generic example of the analytic plan which is prepared by the Task Force as a framework for evaluating the evidence for a screening program. The straight arrows show possible pathways of bene! t, and the blue curved arrows show possible adverse effects relating to different stages. The primary question (question 1 in the ! gure) is generally one of whether screening is effective in reducing the risk of an adverse outcome such as mortality and if so, to what extent.

Generally, few if any studies have examined this overarching question so that the deliberations of the Task Force often deal with the different steps or linkages that comprise this overall pathway. The

TABLE 14-4. U.S. Preventive Services Task Force Levels of Certainty* Regarding Net Benefit

HIGH	The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.
MODERATE	The available evidence is suf! cient to determine the effects of the preventive service on health outcomes, but con! dence in the estimate is constrained by such factors as: • The number, size, or quality of individual studies. • Inconsistency of ! ndings across individual studies. • Limited generalizability of ! ndings to routine primary care practice. • Lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion.
LOW	The available evidence is insuf! cient to assess effects on health outcomes. Evidence is insuf! cient because of: • The limited number or size of studies. • Important ! aws in study design or methods. • Inconsistency of ! ndings across individual studies. • Gaps in the chain of evidence. • Findings not generalizable to routine primary care practice. • A lack of information on important health outcomes. More information may allow an estimation of effects on health outcomes.

*The USPSTF de! nes certainty as "likelihood that the USPSTF assessment of the net bene! t of a preventive service is correct." The net bene! t is de! ned as bene! t minus harm of the preventive service as implemented in a general, primary care population. The USPSTF assigns a certainty level based on the nature of the overall evidence available to assess the net bene! t of a preventive service.

From the U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm/. Accessed August 15, 2013.

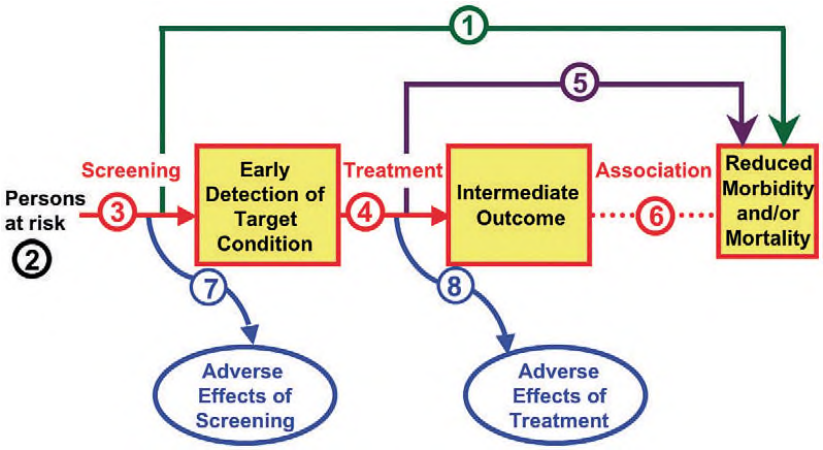


Figure 14-24. Generic analytic framework for screening topics used by the U.S. Preventive Services Task Force. Numbers refer to key questions in the ! gure. (1) Does screening for X reduce morbidity and/or mortality? (2) Can a group at high risk for X be identi- ! ed on clinical grounds? (3) Are there accurate (i.e., sensitive and speci! c) screening tests available? (4) Are treatments available that make a difference in intermediate outcomes when the disease is caught early, or detected by screening? (5) Are treatments available that make a difference in morbidity or mortality when the disease is caught early, or detected by screening? (6) How strong is the association between the intermediate outcomes and patient outcomes? (7) What are the harms of the screening test? (8) What are the harms of the treatment? (Adapted from U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed August 15, 2013.)

purple arrow in the ! gure (step 5) shows the relation of treatment to outcome. Red arrows in the ! gure, steps 3, 4, and 6, show individual components of question 1. These assessments generally depend on a review of relevant randomized trials in order to prepare a chain of supporting evidence on which to base an answer to question 1. The evidence for each linkage is summarized in the evidence review and then summarized across the different linkages to provide an overall assessment of the supporting evidence for the preventive service being evaluated.

The certainty of net bene! t is graded on a 3-point scale: high, moderate, or low (see Table 14-4). The recommendations of the Task Force are based on a combined consideration of the certainty and the magnitude of the net bene! t as shown in the matrix in Figure 14-25, in which a grading system of A, B, C, D, and I is used. The meaning of each letter grade is explained in Table 14-5.

The work of the Task Force has dealt with screening for many diseases and conditions. Some examples will illustrate the breadth of the Task Force's activities. It has reviewed the evidence for screening for different cancers, for cardiovascular diseases including hypertension, coronary heart disease, and abdominal aortic aneurysm, for infectious diseases, such as gonorrhea, Chlamydia, and hepatitis B and C, and for mental conditions such as dementia, depression, and suicide risk, and screening for glaucoma and for type 2 diabetes. The Task Force has

Certainty of Net Benefit	Magnitude of Net Benefit			
	Substantial	Moderate	Small	Zero/ Negative
High	A	B	C	D
Moderate	B	B	C	D
Low	Insufficient Evidence			

Figure 14-25. Grid used by the U.S Preventive Services Task Force for assessing the certainty of bene! t and the magnitude of net bene! t in determining the grade of its recommendations. (Adapted from U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed August 15, 2013.)

also reviewed the evidence for the effectiveness of counseling for many conditions such as counseling to prevent tobacco use and tobacco-related diseases, counseling to prevent alcohol misuse, counseling to promote a healthy diet, and counseling to promote physical activity. The above issues have been addressed in adults, but childhood conditions have also been reviewed by the Task Force including prevention of dental caries in preschool children, screening for scoliosis in adolescents, newborn hearing screening, screening for visual impairment in children younger than 5 years of age, and screening for obesity in children and adolescents. These and many more evidence reviews and recommendations of the Task Force can be found on the website of the Agency for Health Care Research and

TABLE 14-5. **What the USPSTF Grades Mean and Suggestions for Practice**

Grade	Grade Definitions	Suggestions for Practice
A	The USPSTF recommends the service. There is high certainty that the net benefit is substantial.	Offer/provide this service.
B	The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.	Offer/provide this service.
C	The USPSTF recommends against routinely providing the service. There may be considerations that support providing the service in an individual patient. There is at least moderate certainty that the net benefit is small.	Offer/provide this service only if there are other considerations in support of offering/providing the service in an individual patient.
D	The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits.	Discourage the use of this service.
I	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.	Read "Clinical Considerations" section of USPSTF Recommendation Statement. If offered, patients should understand the uncertainty about the balance of benefits and harms.

From the U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed August 15, 2013.

Quality (www.ahrq.gov). The deliberations and recommendations of the Task Force provide a highly useful model of assessing the strength of the evidence and moving from causal inferences to policy recommendations.

CONCLUSION

Although causal guidelines discussed in this chapter are often referred to as *criteria*, this term does not seem entirely appropriate. Although it may be a desirable goal to place causal inferences on a firm quantitative and structural foundation, at present we generally do not have all the information needed for doing so. The preceding list should therefore be considered to be only guidelines that can be of most value when coupled with reasoned judgment about the entire body of available evidence, in making decisions about causation.

In the next chapter, we address several additional issues that need to be considered in deriving causal inferences from epidemiologic studies.

REFERENCES

1. Cameron WB: Informal Sociology: A Casual Introduction to Sociological Thinking. New York, Random House, 1963, p 13. (This quotation was also attributed to Albert Einstein some years later.)
2. Ochsner A, DeBakey M: Primary pulmonary malignancy. Surg Gynecol Obstet 68:435, 1939.
3. Yerushalmy J: Infants with low birth weight born before their mothers started to smoke cigarettes. Am J Obstet Gynecol 112:277–284, 1972.
4. Evans AS: Causation and Disease: A Chronological Journey. New York, Plenum, 1993, pp 13–39.
5. U.S. Department of Health, Education and Welfare: Smoking and Health: Report of the Advisory Committee to the Surgeon General. Washington, DC, Public Health Service, 1964.
6. Schwartz J: Air pollution and daily mortality: A review and meta analysis. Environ Res 64:36–52, 1994.
7. Hill AB: The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine, 58:295–300, 1965.
8. Marshall BJ, Warren JR: Unidentified curved bacilli in the stomachs of patients with gastritis and peptic ulceration. Lancet 1:1311–1315, 1984.
9. Uemura N, Okamoto S, Yamamoto S, et al: *Helicobacter pylori* infection and the development of gastric cancer. N Engl J Med 345:784–789, 2001.
10. Grant BF, Dawson DA: Age at onset of alcohol use and its association with DSM-IV alcohol abuse and dependence: Results from the National Longitudinal Alcohol Epidemiologic Survey. J Subst Abuse 9:103–110, 1997.

11. Merkatz IR, Thompson JE (eds): New Perspectives on Prenatal Care. New York, Elsevier, 1990.
12. Gordis L, Kleinman JC, Klerman LV, et al: Criteria for evaluating evidence regarding the effectiveness of prenatal interventions. In Merkatz IR, Thompson JE (eds): New Perspectives on Prenatal Care. New York, Elsevier, 1990, pp 31–38.
13. U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed August 15, 2013.

REVIEW QUESTIONS FOR CHAPTER 14

1. In a large case-control study of patients with pancreatic cancer, 17% of the patients were found to be diabetic at the time of diagnosis, compared to 4% of a well-matched control group (matched by age, sex, ethnic group, and several other characteristics) that was examined for diabetes at the same time as the cases were diagnosed. It was concluded that the diabetes played a causal role in the pancreatic cancer. This conclusion:
- **a.** Is correct
- **b.** May be incorrect because there is no control or comparison group
- **c.** May be incorrect because of failure to establish the time sequence between onset of the diabetes and diagnosis of pancreatic cancer
- **d.** May be incorrect because of less complete ascertainment of diabetes in the pancreatic cancer cases
- **e.** May be incorrect because of more complete ascertainment of pancreatic cancer in nondiabetic persons

2. An investigator examined cases of fetal death in 27,000 pregnancies and classi! ed mothers according to whether they had experienced sexual intercourse within 1 month before delivery. It was found that 11% of the mothers of fetuses that died and 2.5% of the mothers of fetuses that survived had had sexual intercourse during the period. It was concluded that intercourse during the month preceding delivery caused the fetal deaths. This conclusion:
- **a.** May be incorrect because mothers who had intercourse during the month before childbirth may differ in other important characteristics from those who did not
- **b.** May be incorrect because there is no comparison group
- **c.** May be incorrect because prevalence rates are used where incidence rates are needed
- **d.** May be incorrect because of failure to achieve a high level of statistical signi! cance
- **e.** Both *b* and *c*

3. All of the following are important criteria when making causal inferences *except:*
- **a.** Consistency with existing knowledge
- **b.** Dose-response relationship
- **c.** Consistency of association in several studies
- **d.** Strength of association
- **e.** Predictive value

Questions 4 and 5 are based on the following information.

Factor A, B, or C can each individually cause a certain disease without the other two factors, but only when followed by exposure to factor X. Exposure to factor X alone is not followed by the disease, but the disease never occurs in the absence of exposure to factor X.

4. Factor X is:
- **a.** A necessary and suf! cient cause
- **b.** A necessary, but not suf! cient, cause
- **c.** A suf! cient, but not necessary, cause
- **d.** Neither necessary nor suf! cient
- **e.** None of the above

5. Factor A is:
- **a.** A necessary and suf! cient cause
- **b.** A necessary, but not suf! cient, cause
- **c.** A suf! cient, but not necessary, cause
- **d.** Neither necessary nor suf! cient
- **e.** None of the above

Exhibit 72

Health Canada Santé Canada

Your health and safety… our priority. *Votre santé et votre sécurité… notre priorité.*

Weight of Evidence:
General Principles and Current Applications at Health Canada

PREPARED FOR: Task Force on Scientific Risk Assessment

PREPARED BY: Weight of Evidence Working Group

Canada

Health Canada is the federal department responsible for helping the people of Canada maintain and improve their health. Health Canada is committed to improving the lives of all of Canada's people and to making this country's population among the healthiest in the world as measured by longevity, lifestyle and effective use of the public health care system.

Également disponible en français sous le titre :
Poids de la preuve : Principes généraux et applications actuelles à Santé Canada

To obtain additional information, please contact:

Health Canada
Address Locator 0900C2
Ottawa, ON K1A 0K9
Tel.: 613-957-2991
Toll free: 1-866-225-0709
Fax: 613-941-5366
TTY: 1-800-465-7735
E-mail: hc.publications-publications.sc@canada.ca

Publication date: August 2018

Cat.: H129-69/2018E-PDF
ISBN: 978-0-660-27301-3
Pub.: 180202

Prepared by the Task Force on Scientific Risk Assessment's Weight of Evidence Working Group

T. Tao,[1] Y. Bhuller,[2] Y. Bonvalot,[3] M. Hill,[4] A. Klein,[5] G. Kozak,[6] I. Plante,[7] and N. Robert[8]

Document Revision History

VERSION	DATE	SECTION/PARAGRAPH CHANGED	CHANGE(S) MADE
1	November 2011		Initial issuance of final document.
2	May 2018	Document Revision History	Section added to track revisions.
		Annex 2	Program areas, interpretations and applications updated.
		8.0 References	Section 8.0 added; Reference list updated throughout document.

1 Bioethics and Policy Integration Division, Science Policy Directorate, SPB, Ottawa, ON

2 Health Evaluation Directorate, PMRA, Ottawa, ON

3 Environmental Health Program, Quebec Region, Longueuil, QC

4 New Substances Assessment and Control Bureau, Safe Environments Directorate, HECSB, Ottawa, ON

5 Centre for Evaluation of Radiopharmaceuticals and Biotherapeutics, Biologics and Genetic Therapies Directorate, HPFB, Ottawa, ON

6 Bureau of Microbial Hazards, Food Directorate, HPFB, Ottawa, ON

7 Office of Risk Management and Science, Marketed Health Products Directorate, HPFB, Ottawa, ON

8 Existing Substances Risk Assessment Bureau, Safe Environments Directorate, HECSB, Ottawa, ON

Table of Contents

1. Introduction

Weight of Evidence (WoE) is frequently cited as the basis on which risk assessment conclusions are made. However, multiple interpretations and a lack of consensus about its meaning could potentially compromise communication between diverse stakeholders in the decision-making process. In response to this issue, an analysis of the WoE approach was initiated by Health Canada's Science Policy Directorate in 2010, as a project under the Task Force on Scientific Risk Assessment. By examining current interpretations and identifying potential best practices, this analysis aims to enhance the consistency and coherence of risk assessments across the Department.

2. Purpose and scope

The current document aims to inform senior management about WoE in Health Canada risk assessments by providing an overview of the approach in terms of its:

- role in scientific risk assessments;
- main guiding principles; and
- application by various risk assessment programs at Health Canada.

In addition, this explanatory document serves as a value-added Departmental resource of high level contextual information and guiding principles to supplement program specific guidelines, procedures and/or tools.

While this document acknowledges that WoE could also be applied in the risk management decision making context, where scientific evidence is weighed against other policy considerations, it will not expand on this information as it is considered **not** within the scope of this document.[9]

[9] The terms evidence, information and data are used interchangeably in this document, and refer to general scientific usage, not specific legal definitions of what constitutes evidence, or "admissible" evidence, in a court of law.

3. Role in risk assessments

In general, scientific risk assessments encompass the following steps: identifying and characterizing the hazard, assessing the exposure, and characterizing the risk; risk assessments also play an integrated role in an evidence- informed decision making process which also involves managing and communicating the risk.

WoE in the risk assessment context is defined in *Health Canada Decision-Making Framework for Identifying, Assessing, and Managing Health Risks* (Health Canada, 2000) as:

> *"A qualitative measure that takes into account the nature and quality of scientific studies intended to examine the risk of an agent. Uncertainties that result from the incompleteness and unavailability of scientific data frequently require scientists to make inferences, assumptions, and judgements in order to characterize a risk. Making judgements about risk based on scientific information is called "evaluating the weight of evidence".*

The above description can be interpreted to implicitly include two separate concepts frequently associated with WoE terminology:

1. **Totality of Evidence:** what types and sources of information are to be gathered and considered for subsequent assessment; and
2. **Weighing Evidence:** how such individual sources of evidence are assessed and integrated into an overall conclusion or recommendation.

Totality of evidence can be influenced by varying interpretations of *"all"* available or relevant evidence to date. This concept provides the opportunity to make use of information/studies that may be regarded insufficient individually, but which contribute to a total "weight of evidence" case in support of conclusions during risk assessment when they are considered alongside other studies/sources of evidence. Moreover, an evaluation of evidence and of any subsequent decision can be reassessed, at a later date, based on the availability of data that may not have been readily available at the time of the original assessment.

The latter, methodological concept of weighing evidence is applicable to most risk assessments. While specific methodologies and tools used for assessing and integrating evidence (e.g., quantitative or qualitative) may vary and are context dependent, the general principles for the assessment and integration process remain the same.

4. General principles

The inter-relationship of the above two concepts, and the general principles of the WoE approach outlined below, is presented in Annex 1, for illustrative purposes only. The Totality of Evidence concept includes the general principles 4.1, 4.2, and 4.3, while the methodological concept of Weighing Evidence can be subdivided into the general principles of 4.4 and 4.5. Regardless of specific interpretations of terminology, the following steps are applicable to building a "weight of evidence" case for a given risk assessment conclusion or recommendation.

4.1 GATHERING *"ALL"* AVAILABLE EVIDENCE

Multiple sources and types of evidence may be gathered or submitted and considered in context of *"all"* available evidence to date. Depending on the regulatory data requirements, the full spectrum of sources and types of evidence may include: randomized controlled clinical trials, company and/or third party generated studies of a proprietary nature, peer-reviewed, published scientific literature, expert opinion reports, decisions and analysis reports from regulatory authorities, incident reports, adverse reactions submitted to regulatory authorities, and unpublished data.

4.2 ASSESSXIVIDUAL STUDIES

General criteria for inclusion/exclusion are useful when screening *"all"* evidence gathered for further consideration. While specific terminology and scope for inter-related screening criteria such as "quality", "reliability", "relevance", etc. could differ across various regulatory programs and agencies, the underlying principles are common. Assessment could involve use of specific scoring tools and/or best professional judgement. Acceptable studies that meet standards for inclusion are assessed further in subsequent steps of the WoE approach, while unacceptable studies may be excluded from further consideration. For example, unpublished data, or data irrelevant to the risk assessment endpoint in question, may be excluded from further consideration or may be given a lower weight when assembling the lines of evidence. When necessary, the rationale for including (or excluding) studies could be documented in the relevant report.

4.3 ASSEMBLING LINES OF EVIDENCE

The types and sources of evidence considered are diverse and vary considerably in level of detail. Depending on the context of the risk assessment in question, individual studies or data sources are often assessed as distinct lines of evidence on their own, or considered in concert with other similar studies that together constitute a particular "line of evidence". Such lines can be organized according to unifying characteristics, such as source or type of data (e.g., animal data, human data, clinical trials, and literature data). Separate lines of evidence can also be drawn along sub- components of risk, such as hazard, exposure, human health, environmental safety, or other characteristics such as studies which support or counter a particular conclusion. These lines can be further subdivided into more specific lines. For example, "hazard" can be divided into specific organ systems (hazard to the liver, kidneys, brain, etc.).

4.4 ASSESSING LINES OF EVIDENCE

Lines of evidence are assessed against various criteria that are dependent on the context of the particular endpoint in question. Risk assessments can be hypothesis driven, and designed to answer yes/no questions (e.g., is substance *x* a carcinogen?). In such instances, several lines of evidence (e.g., carcinogenicity studies, genotoxicity studies, or mechanistic data) can each be assessed based on criteria such as the strength/robustness of evidence in support of, or against, a given conclusion for each particular line.

Other risk assessments can address more general questions (e.g., what product/source is the likely cause of illness outbreak *y*?). In such instances, some lines of evidence, such as epidemiological data, can be assessed based on specific criteria such as strength of association, consistency, specificity, temporality, biological gradient/ dose-response, plausibility, coherence, experimental evidence, and analogy (e.g., the Bradford Hill (1965) criteria for causal inference). Depending on the context of the particular line of evidence involved, other criteria not described here could also be applicable. The assessment can be quantitative, by assigning a weight or value to each line of evidence assessed, in the form of probabilities, alphanumeric values, or qualitative by descriptions such as "weak" or "strong", or implicit, in the form of logic models and decision trees that by default emphasize the importance of certain lines of evidence over others.

Assigned values or descriptions reflect the relative "strength" of a particular line of evidence, which is negatively impacted by the uncertainty and variability in datasets contributing to each line of evidence. Departmental documents elaborating on uncertainty and/or variability include, but are not limited to: *A Framework for the Application of Precaution in Science-based Decision Making about Risk* (Privy Council Office, 2003) and the *Health Products and Food Branch's Guide for Conducting Health Risk Assessments in Humans* (Health Canada, 2011).

4.5 INTEGRATING MULTIPLE LINES OF EVIDENCE

The determination of the relative contributions of various lines of evidence to the overall conclusion can be performed in a single step, qualitative process, using best professional judgment. More systematic methods of quantitative integration can also be employed, where scores for individual lines of evidence may be adjusted by weighting factors that reflect the relative importance of a line within the overall body of evidence, and then mathematically integrated into a final value. However, scoring is not easily applicable in a context such as risk assessment, due to the large complexity of the different sources of information available.

The integration of values/weights is an iterative process that is repeated at many levels: within individual studies, across similar studies into a collective value for a particular line of evidence, and across multiple lines of evidence into an overall risk assessment conclusion or recommendation. For example, to determine whether a compound affects the liver, one collectively examines and integrates clinical chemistry findings along with organ weight and histopathology data within a single study, or across multiple similar studies (e.g., to assess dose-response). For integration across collection of studies for a given assessment endpoint (e.g., whether a compound is carcinogenic) one can collectively examine and integrate carcinogenicity studies, genotoxicity studies and mechanistic data. For conclusions regarding overall risk, it is necessary to integrate lines of evidence related to hazard and exposure. Further integration of human health risk and environmental risk may contribute to an overall risk profile.

5. Application at Health Canada

Assessment of scientific evidence is a crucial component of risk assessment and decision making at Health Canada. Moreover, for many of the regulatory programs in the Department, risk assessment conclusions (referred to as risk characterization) are often made based on the likelihood of association between a particular substance/activity and associated health effects. In this context, a WoE approach is frequently cited as the basis on which conclusions are made using the best available information to date that can be gathered, assessed, and integrated using various qualitative and quantitative methods.

The mandate and scope of risk assessment and/or risk management activities of the various programs vary significantly across the Department (see *A Primer on Scientific Risk Assessment at Health Canada,* Saner, 2010). Each program operates within the constraints of program-specific legislation. Differences in legislation and program goals impact time available for assessment of each particular product or activity, the amount and quality of information that is available to date for assessment, and

the degree of flexibility in interpretation and application of WoE as a risk assessment approach. Each program is also impacted by international guidelines for specific subject areas and the sector-specific context in which regulations may be often harmonized globally. The varying issues and the context under which regulatory decisions are made, and the scope of potential risk management options and recommendations that can be explored also differ across and within programs.

A survey was conducted to determine how WoE was applied across the department. All branches surveyed responded, including the Health Products and Food Branch (HPFB), the Healthy Environments and Consumer Safety Branch (HECSB), and the Pesticide Management Regulatory Agency (PMRA). The general principles of the WoE approach are applied by most programs surveyed. Specifically, most risk assessments follow the steps of gathering and assessing individual studies and sources of evidence, assembling studies into context specific lines of evidence, and assessing and integrating multiple lines of evidence into an overall conclusion or recommendation. Most programs interpret WoE to include concepts such as the totality of evidence (i.e., the evidence to be gathered and considered), as well as the weighing of evidence (i.e., how such evidence is assessed and integrated into a final conclusion) (see Annex 2).

The application of specific criteria and tools are context specific, and are outlined in various program specific guidelines, standard operating procedures, working documents, etc. Program documents outlining application of the WoE approach have been specifically developed for such purposes when considered necessary. For example:

- *Weight of Evidence: Factors to Consider for Appropriate and Timely Action in a Foodborne Illness Outbreak Investigation* (Health Canada, Public Health Agency of Canada, and Canadian Food Inspection Agency, 2011);
- *Framework for Initiating and Conducting Risk Analysis Activities on Microbial Hazards in Food* (Health Canada, 2017);
- *Food Investigation Response Manual* (Canadian Food Inspection Agency, 2017);
- *Science Policy Note: General Exposure Factor Inputs for Dietary, Occupational, and Residential Exposure Assessments* (Health Canada, 2014)
- *Federal Contaminated Site Risk Assessment in Canada: Supplemental Guidance on Human Health Risk Assessment for Country Foods (HHRAFoods)* (Health Canada, 2010);
- *Notice to Product License Applicants—Traditional Claim Submissions: Evidence Criteria and Evidence Assessment Template* (Natural and Non-prescription Health Products Directorate, 2010);

- *Pathway for Licensing Natural Health Products used as Traditional Medicines* (Natural and Non-prescription Health Products Directorate, 2012a)
- *Pathway for Licensing Natural Health Products Making Modern Health Claims* (Natural and Non-prescription Health Products Directorate, 2012b)

Program documents of a more general nature include:

- *Health Products and Food Branch's Guide for Conducting Health Risk Assessments in Humans* (Health Canada, 2011);
- *Framework for Science-Based Risk Assessment of Micro-Organisms Regulated under the Canadian Environmental Protection Act, 1999* (Environment Canada, Health Canada, 2013);
- *All Hazards Risk Assessment Methodology Guidelines 2012–2013* (Public Safety Canada 2018)

As mentioned above, documentation on how the risk assessment is conducted and the rationale for either including or excluding certain sources of evidence is a critical component of the decision making process. Similarly, while the WoE approach is consistently applied in most risk assessments across the Department, explicit use of WoE terminology is not always documented. In some instances, WoE terminology is used, but the specific application of the WoE approach is not elaborated.

On occasion, WoE terminology is used when actually referring to levels of evidence or standards of quality of individual studies. In some instances, WoE terminology is also used in place of actual descriptions of the strength/robustness of overall conclusions/recommendations, or in place of legal terms such as preponderance of evidence, which simply means more likely than not.

The majority of risk assessment reports, however, provide a logical narrative description of the relative strengths or weaknesses of various lines of evidence considered. For most risk assessments, individual lines of evidence are pooled and integrated into a final conclusion based on best professional judgment, and not mathematical formula. Narrative descriptions of the rationale for such judgments are usually provided, including explanations of how certain lines of evidence are more important than others in determining the overall risk assessment conclusion/recommendation. Some reports, however, simply list lines of evidence assessed and proceed directly to the overall risk assessment conclusion, without explicit documentation of how the multiple lines of evidence relate to one another, or the rationale behind the integration process.

6. International context

The WoE approach is routinely applied by most scientific risk assessment agencies internationally and while several definitions for WoE exist, there is no single, universal standardized/commonly agreed upon definition or specific guidance on how to implement a WoE approach. For example, recent guidance on the use of the WoE approach and "totality of evidence" has been published by the European Food Safety Authority's Scientific Committee (EFSA, 2017a), which stated that "*weight of evidence assessment is a process in which evidence is integrated to determine the relative support for possible answers to a scientific question. The term 'weight of evidence' on its own is the extent to which evidence supports possible answers to a scientific question.*"

The United States Environmental Protection Agency (EPA, 2003) outlines WoE in various guidelines, in both the totality of evidence context, and the methodological context of the weighing of multiple lines of evidence, e.g.:

> *"The weight-of-evidence approach considers all relevant information in an integrative assessment that takes into account the kinds of evidence available, the quality and quantity of the evidence, the strengths and limitations associated with each type of evidence and explains how the various types of evidence fit together."*

However, in a review of the EPA's Integrated Risk Information System (IRIS) process, the National Research Council (2014) found that:

> *"systematic review and weight-of-evidence analysis have historically been described in various ways, and the terms are sometimes used interchangeably; this vagueness in use of terminology results in some confusion as to what the terms mean in practice… The committee views weight-of-evidence analysis as a judgment-based process for evaluating the strength of evidence to infer causation. However, it found that the phrase as used in practice has become too vague and is of little scientific use. An IRIS assessment must come to a judgment about whether a chemical is hazardous to human health and must do so by integrating a variety of lines of evidence. Therefore, the committee found the term evidence integration to be more useful and more descriptive of the process that occurs after completion of systematic reviews."*

Similarly, the U.S. Environmental Protection Agency's National Center for Environmental Assessment (2015) takes an integrated approach to science assessments for reviews of national ambient air quality standards:

"The U.S. EPA integrates the evidence from across scientific disciplines or study types and characterizes the weight of evidence for relationships… drawing upon the results of all studies judged of adequate quality and relevance per the criteria… consider aspects, such as strength, consistency, coherence, and biological plausibility of the evidence, and develop causality determinations on the nature of the relationships… includes evaluating strengths and weaknesses in the overall collection of studies across disciplines."

The European Chemicals Agency (ECHA, 2011 and 2016) outlines interpretations regarding the methodological context of weighing evidence as follows:

"The weight of evidence approach commonly refers to combining evidence from multiple sources to assess a property under consideration. It can therefore be a useful technique where, for example, each piece of information or test alone is not sufficient to address a standard information requirement but where it may be possible to combine the strengths and weaknesses of the individual studies to reach a conclusion for a particular property.

The term weight of evidence (WoE) is neither a scientifically well-defined term nor an agreed formalised concept characterised by defined tools and procedures. It can, however, be regarded as an evidence-based approach involving an assessment of the relative weights (values) of different pieces of the available information that have been gathered. Application of this concept can be achieved either in an objective way by using a formalised procedure or by using expert judgement. Factors such as the quality of the data, consistency of results, nature and severity of effects, relevance of the information will have an influence on the weight given to the available evidence."

This concept of weighing evidence is supplemented by the totality of evidence concept within the Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (ECHA, 2017):

"There may be sufficient weight of evidence from several independent sources of information leading to the assumption/conclusion that a substance has or has not a particular dangerous property, while the information from each single source alone is regarded insufficient to support this notion.

There may be sufficient weight of evidence from the use of newly developed test methods, not yet included in the test methods referred to in Article 13(3) or from an international test method recognised by the Commission or the Agency as being equivalent, leading to the conclusion that a substance has or has not a particular dangerous property.

Where sufficient weight of evidence for the presence or absence of a particular dangerous property is available:
— further testing on vertebrate animals for that property shall be omitted,
— further testing not involving vertebrate animals may be omitted.

In all cases adequate and reliable documentation shall be provided."

The World Health Organization's International Programme on Chemical Safety has published two guidance documents regarding uncertainty in risk assessment: *Uncertainty and Data Quality in Exposure Assessment* (WHO, 2008), which explicitly addresses WoE: *"to the extent possible, the combined effect of different sources of uncertainty on the exposure or risk predictions, perhaps based on a weight-of-evidence methodology in the absence of quantitative data, should also be considered"*, and a *Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization* (WHO, 2017).

The Food and Agriculture Organization of the United Nations and the World Health Organization (FAO/ WHO, 2009) discussed using a WoE approach to the risk characterization of microbiological hazards in food: *"the weight of evidence should be evaluated according to clearly specified, scientific criteria. As more criteria are satisfied, the weight of evidence indicates a more credible risk."* FAO/WHO anticipated that *"weight-of-evidence determinations will become increasingly prominent in risk assessments of microbiological pathogens in food."*

The Organisation for Economic Co-operation and Development (OECD, 2015, 2017, 2018) defines WoE as *"a comprehensive, integrated, often qualitative judgment of the extent and quality of information supporting an hypothesis for which the approaches and tools vary, depending on the context."* WoE methodology is used in their "Adverse Outcome Pathway (AOP)/Mode Of Action (MOA)" framework for the development and use of "Integrated Approaches to Testing and Assessment" (IATA):

> *"Evaluation of existing information or generation of additional data within an IATA can be performed on the basis of a non-formalised Weight of Evidence (WoE) approach or by using predefined, structured approaches such as Sequential Testing Strategies (STS), Integrated Testing Strategies (ITS) or their combination."*

In considering the use of a WoE approach, Codex Alimentarius (2014) cautions that *"The weight of evidence integrating quantitative and qualitative data may permit only a qualitative estimate of risk."*

Taken together, the above definitions from key international partners are consistent with current Health Canada interpretations of the WoE approach.

7. Conclusions

While specific tools and methodologies are often context-specific to particular program areas, the underlying principles of the WoE approach, in which multiple sources of information are gathered, assessed, and integrated into an overall conclusion, are commonly applied across the Department, and are judged to be consistent with international practice.

Presently, inconsistencies occur not in the high level applications of the overall WoE approach. Rather, they result when WoE terminology is applied when actually dealing with standards of quality of individual studies or strength of overall conclusions/recommendations.

Given the context specific nature of each risk assessment and the diversity of tools and criteria applicable, transparent documentation of the specific application of the WoE approach is especially important. There are opportunities for harmonization, and adherence to a simple checklist is a step towards this goal (see Annex 3). Program areas are encouraged to take the relevant steps (e.g., updating internal guidelines) to further improve the documentation aspect in reports that provide the risk assessment in support of subsequent risk management options/regulatory decision, which includes elaborating on what is meant by WoE, when necessary. Additionally, graphically based evidence maps, profiles, or tables may be helpful as supplementary tools for communication from risk assessors to risk managers.

8. References

Bradford Hill A. 1965. The environment and disease: association or causation? *Proc R Soc Med.* 58(5): 295–300.

Canadian Food Inspection Agency. 2017. *Food Investigation Response Manual.* Available at: www.inspection.gc.ca/food/safe-food-production-systems/food-recall-and-emergency-response/food-manual/eng/1378402475724/1378403080658?chap=0 (accessed 2018-03-16)

Canada Privy Council Office. 2003. *A Framework for the Application of Precaution in Science-based Decision Making about Risk*. Available at: http://publications.gc.ca/collections/Collection/CP22-70-2003E.pdf (accessed 2018-03-15)

Codex Alimentarius. 2014. *Principles and Guidelines for the Conduct of Microbiological Risk Assessment (CAC/GL 30-1999)*. Available at: www.fao.org/fao-who-codexalimentarius/codex-texts/guidelines/en/ (accessed 2018-03-20)

ECHA: European Chemicals Agency. 2011. *Guidance on information requirements and chemical safety assessment—Chapter R.4: Evaluation of available information*. Available at: https://echa.europa.eu/documents/10162/13643/information_requirements_r4_en.pdf/d6395ad2-1596-4708-ba86-0136686d205e (accessed 2018-03-19).

ECHA. 2016. *Practical guide: How to use alternatives to animal testing to fulfil your information requirements for REACH registration.* Available at: https://echa.europa.eu/documents/10162/13655/practical_guide_how_to_use_alternatives_en.pdf/148b30c7-c186-463c-a898-522a888a4404 (accessed 2018-03-19)

ECHA. 2017. *Consolidated version of the REACH Regulation: Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)*. Available at: https://echa.europa.eu/regulations/reach/legislation (accessed 2018-03-19)

Environment Canada, Health Canada. 2013. *Framework for Science-Based Risk Assessment of Micro-Organisms Regulated under the* Canadian Environmental Protection Act, 1999. Available at: www.ec.gc.ca/subsnouvelles-newsubs/default.asp?lang=En&n=120842D5-1 (accessed 2018-03-16))

EFSA: European Food Safety Authority. 2017a. *Guidance on the use of the weight of evidence approach in scientific assessments*. EFSA Journal 2017;15(8):4971. Available at: http://onlinelibrary.wiley.com/doi/10.2903/j.efsa.2017.4971/epdf (accessed 2018-03-15)

EFSA. 2017b. *Revised Draft for Internal Testing: Guidance on Uncertainty in EFSA Scientific Assessment*. Available at: www.efsa.europa.eu/sites/default/files/160321DraftGDUncertaintyInScientificAssessment.pdf (accessed 2018-03-15)

EPA: United States Environmental Protection Agency. 2003. *A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information*. Available at: www.epa.gov/sites/production/files/2015-01/documents/assess2.pdf (accessed 2018-03-19)

EPA: United States Environmental Protection Agency. 2015. Preamble to the Integrated Science Assessments (ISA). Available at: https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310244 (accessed 2018-03-29)

FAO/WHO: Food and Agriculture Organization of the United Nations and the World Health Organization. 2009. *Risk characterization of microbiological hazards in food: Guidelines. Microbiological Risk Assessment Series 17.* Available at: www.fao.org/docrep/012/i1134e/i1134e.pdf (accessed 2018-03-20)

Health Canada. 2000. *Health Canada Decision-Making Framework for Identifying, Assessing, and Managing Health Risks.* Available at: www.canada.ca/content/dam/hc-sc/migration/hc-sc/ahc-asc/alt_formats/hpfb-dgpsa/pdf/pubs/risk-risques-eng.pdf (accessed 2018-03-15)

Health Canada. 2010. *Federal Contaminated Site Risk Assessment in Canada: Supplemental Guidance on Human Health Risk Assessment for Country Foods (HHRAFoods).* Available at: http://publications.gc.ca/collections/collection_2012/sc-hc/H128-1-11-641-eng.pdf (accessed 2018-03-16)

Health Canada. 2011. *Health Products and Food Branch's Guide for Conducting Health Risk Assessments in Humans*. Available at: http://mysource.hc-sc.gc.ca/sites/default/files/hpfbs_guide_for_conducting_hras_in_humans_e_final.pdf (accessed 2018-03-15)

Health Canada. 2014. *Science Policy Note: General Exposure Factor Inputs for Dietary, Occupational, and Residential Exposure Assessments*. Available at: www.canada.ca/content/dam/hc-sc/migration/hc-sc/cps-spc/alt_formats/pdf/pubs/pest/pol-guide/spn2014-01/spn2014-01-eng.pdf (accessed 2018-03-16)

Health Canada. 2017. *Framework for Initiating and Conducting Risk Analysis Activities on Microbial Hazards in Food.* Cat.: H164-198/2017E-PDF, ISBN: 978-0-660-06811-4, Pub.: 160231. Available from publications@hc-sc.gc.ca

Health Canada, Public Health Agency of Canada, Canadian Food Inspection Agency. 2011. *Weight of Evidence: Factors to Consider for Appropriate and Timely Action in a Foodborne Illness Outbreak Investigation*. Available at: www.canada.ca/content/dam/hc-sc/migration/hc-sc/fn-an/alt_formats/pdf/pubs/securit/2011-food-illness-outbreak-eclosion-malad-ailments-eng.pdf (accessed 2018-03-16)

National Research Council. 2014. *Review of EPA's Integrated Risk Information System (IRIS) Process*. Committee to Review the IRIS Process; Board on Environmental Studies and Toxicology; Division on Earth and Life Studies. National Academies of Sciences, Engineering, Medicine – National Academies Press. Available at: http://nap.edu/18764 (accessed 2018-03-19)

Natural and Non-prescription Health Products Directorate, Health Canada. 2010. *Notice to Product License Applicants – Traditional Claim Submissions: Evidence Criteria and Evidence Assessment Template.* Available at: www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/legislation-guidelines/guidance-documents/traditional-claim-submissions-evidence-criteria-evidence-assessment-template.html (accessed 2018-03-16)

Natural and Non-prescription Health Products Directorate, Health Canada. 2012a. *Pathway for Licensing Natural Health Products used as Traditional Medicines.* Available at: www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/legislation-guidelines/guidance-documents/pathway-licensing-traditional-medicines.html (accessed 2018-03-16)

Natural and Non-prescription Health Products Directorate, Health Canada. 2012b. *Pathway for Licensing Natural Health Products Making Modern Health Claims*. Available at: www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/legislation-guidelines/guidance-documents/pathway-licensing-making-modern-health-claims.html (accessed 2018-03-16)

OECD: Organisation for Economic Co-operation and Development. 2015. *Report of the Workshop on a Framework for the Development and Use of Integrated Approaches to Testing and Assessment. Series on Testing and Assessment No. 215, ENV/JM/MONO(2015)22*. Available at: www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2015)22&doclanguage=en (accessed 2018-03-20)

OECD. 2017. *Revised Guidance Document on Developing and Assessing Adverse Outcome Pathways. Series on Testing & Assessment No. 184, ENV/JM/MONO(2013)6.* Available at: www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2013)6&doclanguage=en (accessed 2018-03-20)

OECD. 2018. *Users' Handbook Supplement to the Guidance Document for Developing and Assessing AOPs. Series on Testing & Assessment No. 233, Series on Adverse Outcome Pathways No. 1, ENV/JM/MONO(2016)12*. Available at: https://one.oecd.org/document/ENV/JM/MONO(2016)12/en/pdf (accessed 2018-03-20)

Public Safety Canada. 2018. *All Hazards Risk Assessment Methodology Guidelines 2012-2013*. Available at: www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ll-hzrds-ssssmnt/index-en.aspx (accessed 2018-03-16)

Saner M. 2010. *A Primer on Scientific Risk Assessment at Health Canada.* HC Pub.: 100140; Cat.: H22-4/3-2010; ISBN: 978-1-100-15377-3. Available at: www.canada.ca/content/dam/hc-sc/migration/hc-sc/sr-sr/alt_formats/pdf/pubs/about-apropos/2010-scientif-ris-eng.pdf (accessed 2018-03-16)

WHO (World Health Organization). 2008. *International Programme on Chemical Safety: Uncertainty and Data Quality in Exposure Assessment.* Available at: www.who.int/ipcs/publications/methods/harmonization/exposure_assessment.pdf?ua=1 (accessed 2018-03-15)

WHO. 2017. *International Programme on Chemical Safety: Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization.* Available at: http://apps.who.int/iris/bitstream/10665/259858/1/9789241513548-eng.pdf?ua=1 (accessed 2018-03-15)

Annex 1

Graphical Representation of the Weight of Evidence Approach in the Context of Scientific Risk Assessments

Totality of Evidence: ***conceptual interpretations of scope and nature of evidence for consideration***

1. GATHERING "ALL" EVIDENCE

2. ASSESSING INDIVIDUAL STUDIES (for 'quality', 'relevance', etc…)

(Note: Individual studies may not meet standards for all criteria, but still could be considered if contributing to a weight of evidence case when pooled with other studies.)

3. ASSEMBLING LINES OF EVIDENCE (context specific)

Weighing Evidence: ***methodological approaches for assessing and integrating evidence into an overall conclusion***

4. ASSESSING LINES OF EVIDENCE (for causality, hazard, exposure, etc…)
5. INTEGRATING MULTIPLE LINES OF EVIDENCE

Example 1: Weight of Evidence Summary for Foodborne Illness Outbreak[1]

Lines of Evidence	Criteria assessed	Qualitative Description
Line 1 (Laboratory Investigation: Strength of Microbiological Evidence: Determining the Relevance of Food and Clinical Isolates that Match by Pulsed-Field Gel Electrophoresis (PFGE)*)	A. Does the organism show suitable diversity by PFGE? B. Are clinical and food isolate PFGE patterns indistinguishable by enzymes? C. What is the historic frequency of the PFGE pattern combination? D. Are other subtyping results available; are they consistent with PFGE?	(Weak….Strong)
Line 2 (Food Safety Investigation: Traceback and Trace Forward)	• Does the product identity allow direct tracing to the manufacturer? • Can the point of purchase be identified? • Can the product identity lead to manufacturer? • Distribution Channel • Can the product be traced back to the wholesaler?	(Weak….Strong)
Line 3 (Epidemiological Investigation)	• Plausibility, • Consistency, • Specificity, • Strength of the association, • Temporal(ity), • Dose-Response, • etc…	(Weak….Strong)
Overall condusion	The WoE case for product Y as source of contamination for Foodborne Illness Outbreak X is (Weak….Strong)	

Example 2: Weight of Evidence Summary for neurobehavioural effects of bisphenol-A at low doses in rodents (below reproductive/ developmental NOAEL)[2]

Criteria Assessed	Qualitative Description
Rigour of overall dataset	limited
Power of overall dataset	limited
Corroboration/consistency	limited
Biological Plausibility/coherence	limited
Overall weight of evidence for neurodevelopmental effects in rodents at low level exposures (exposure below the NOAEL for reproductive/developmental toxicity)	limited

1 *Weight of Evidence: Factors to Consider for Appropriate and Timely Action in a Foodborne Illness Outbreak Investigation. Health Canada, Public Health Agency of Canada, and Canadian Food Inspection Agency. January 2011*

2 *Final Screening Assessment for the Phenol, 4,4' -(1-methylethylidene)bis-(Bisphenol A). Chemical Abstracts Service Registry Number 80-05-7. Environment Canada and Health Canada. October 2008.*

Annex 2

Working level interpretation(s)/application(s) of Weight of Evidence

PROGRAM	*TOTALITY OF EVIDENCE* *: Conceptual interpretations of the nature and scope of evidence sources for consideration		*WEIGHING EVIDENCE* **: Methodological approaches for the assessment and integration of multiple lines of evidence to derive at a final conclusion		
	Consideration of all available lines of evidence to date, as opposed to a subset of data	Consideration of studies that individually may not meet standards for all criteria, but contributing to a weight of evidence case when pooled with other studies	Qualitative (e.g., listing, best professional judgment)	Semi-quantitative (e.g., causal criteria, logic models, alphanumeric scoring or indexing)	Quantitative (e.g., probabilistic tools or Multi-Criteria Decision Analysis [MCDA])
Healthy Environments and Consumer Safety Branch (HECSB)					
WAQB[1] / Water	✓	✓	✓	✓	
WAQB / Air	✓	✓	✓	✓	
New Substances – NSACB[2]	✓	✓	✓		
Existing Substances – ESRAB[3]	✓	✓	✓	✓	
ERHSD[4]			✓	✓	
CPSD[5]	✓	✓	✓		✓
Health Products and Food Branch (HPFB)					
TPD[6]	✓	✓	✓	✓	✓
BGTD[7] / Biologics	✓	✓	✓	✓	✓
MHPD[8]	✓	✓	✓	✓	✓
NNHPD[9]	✓	✓	✓	✓	
VDD[10]	✓	✓	✓	✓	✓
FD[11] / Novel Foods	✓	✓	✓		
FD / Nutrition Labelling and Claims		✓			
FD / Microbial Hazards	✓	✓	✓	✓	✓
Pesticide Management Regulatory Agency (PMRA)					
HED[12]	✓	✓	✓	✓	✓

1 Water and Air Quality Bureau, Safe Environments Directorate (SED)

2 New Substances Assessment and Control Bureau, SED

[3] Existing Substances Risk Assessment Bureau, SED

[4] Environmental and Radiation Health Sciences Directorate

[5] Consumer Product Safety Directorate

[6] Therapeutic Products Directorate

[7] Biologics and Genetic Therapies Directorate

[8] Marketed Health Products Directorate

[9] Natural and Non-prescription Health Products Directorate

[10] Veterinary Drugs Directorate

[11] Food Directorate

[12] Health Evaluation Directorate

* In general, the totality of evidence concept does not involve any actual "weighing" of multiple lines of evidence relative to each other, and is thus not interpreted as part of the WoE concept by certain programs. Nevertheless, this concept is commonly recognized as part of the scope of the WoE approach in the risk assessment context by most programs across the Department. Some differences are also observed regarding the sub-concept of considering "all" available evidence and such apparent differences may be the result of more literal interpretations of "all" available evidence by these programs compared to others, rather than a true reflection of actual differences of risk assessment practices. Moreover, a precise interpretation of "all" is also dependent on the program area/regulatory requirements in terms of the type of evidence that are required in order to support a submission. For example, for programs conducting risk assessments on therapeutic products such as drugs and biologics, the requirements come from guidelines of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). On the other hand, pesticide evaluations utilize guidance lists of required and conditionally required data, which differ depending on how and where the product is used.

** Qualitative methods of assessing and integrating multiple lines of scientific evidence in Departmental risk assessment programs seem to be the dominant application of the WoE approach for risk assessments across the Department. Specific qualitative methods can range from simple listing of the evidence assessed, to more detailed narrative descriptions that explain the rationale behind the application of best professional judgment, in which some lines of evidence are considered more important, and are given more weight, compared to others. Semi-quantitative methods include systematic assignment of alphanumeric scores for each line of evidence, as well as logic models, decision trees and causality analysis that implicitly give more weight to certain lines of evidence over others, even if actual numeric scores are not assigned. Frequently referred to as "levels of evidence", scoring tools are more often applied by programs involved in the regulation of therapeutic products, for which such standards and scoring systems exist, and are practiced, by international counterparts. Semi-quantitative methods employing hierarchal descriptors instead of alphanumeric scores are frequently employed in the context of assigning value to causality criteria used in foodborne illness outbreaks. Similarly, semi-quantitative descriptors for key events could be used for Mode of Action (MoA) assessment of chemicals for carcinogenicity/mutagenicity. Quantitative methods are not widely used across the Department, often due to limitations in availability of appropriate data.

Annex 3

Checklist for Transparent Documentation of Weight of Evidence Approach

When weight of evidence terminology is used, specify intended meaning in relation to the following concepts:

- ☐ ***Totality of Evidence:*** conceptual interpretations of the nature and scope of evidence sources for consideration
- ☐ ***Weighing Evidence:*** methodology for assessment and integration of multiple lines of evidence

For the risk assessment process, are the following documented?

- ☐ evidence gathered: all available to date, individual sources and types
- ☐ evidence included for further consideration, and why (i.e., inclusion criteria)
- ☐ evidence excluded from further consideration, and why (i.e., exclusion criteria)
- ☐ lines of evidence assembled (list individual studies under each line)
- ☐ assessment criteria applied to lines of evidence, and scoring tools used (if any)
- ☐ values/weighting assigned to each line of evidence (e.g., descriptions, alphanumeric)
- ☐ integration scheme (e.g., best professional judgment, mathematical formula, criteria framework)
- ☐ overall conclusion/recommendation(s)

Exhibit 73

Lesson 1: Introduction to Epidemiology

Section 1: Definition of Epidemiology

The word epidemiology comes from the Greek words *epi*, meaning on or upon, *demos*, meaning people, and *logos*, meaning the study of. In other words, the word epidemiology has its roots in the study of what befalls a population. Many definitions have been proposed, but the following definition captures the underlying principles and public health spirit of epidemiology:

*Epidemiology is the **study** of the **distribution** and **determinants** of **health-related states or events** in **specified populations**, and the **application** of this study to the control of health problems* (*1* (http://www.cdc.gov/OPHSS/CSELS/DSEPD/SS1978/Lesson1/Section1.html#_ref1)).

Key terms in this definition reflect some of the important principles of epidemiology.

Study

Epidemiology is a scientific discipline with sound methods of scientific inquiry at its foundation. Epidemiology is data-driven and relies on a systematic and unbiased approach to the collection, analysis, and interpretation of data. Basic epidemiologic methods tend to rely on careful observation and use of valid comparison groups to assess whether what was observed, such as the number of cases of disease in a particular area during a particular time period or the frequency of an exposure among persons with disease, differs from what might be expected. However, epidemiology also draws on methods from other scientific fields, including biostatistics and informatics, with biologic, economic, social, and behavioral sciences.

Students of journalism are taught that a good news story, whether it be about a bank robbery, dramatic rescue, or presidential candidate's speech, must include the 5 W's: what, who, where, when and why (sometimes cited as why/how). The 5 W's are the essential components of a news story because if any of the five are missing, the story is incomplete.

The same is true in characterizing epidemiologic events, whether it be an outbreak of norovirus among cruise ship passengers or the use of mammograms to detect early breast cancer. The difference is that epidemiologists tend to use synonyms for the 5 W's: diagnosis or health event (what), person (who), place (where), time (when), and causes, risk factors, and modes of transmission (why/how).

In fact, epidemiology is often described as the basic science of public health, and for good reason. First, epidemiology is a quantitative discipline that relies on a working knowledge of probability, statistics, and sound research methods. Second, epidemiology is a method of causal reasoning based on developing and testing hypotheses grounded in such scientific fields as biology, behavioral sciences, physics, and ergonomics to explain health-related behaviors, states, and events. However, epidemiology is not just a research activity but an integral component of public health, providing the foundation for directing practical and appropriate public health action based on this science and causal reasoning.(*2* (http://www.cdc.gov/OPHSS/CSELS/DSEPD/SS1978/Lesson1/Section1.html#_ref2))

Distribution

Epidemiology is concerned with the **frequency** and **pattern** of health events in a population:

> **Frequency** refers not only to the number of health events such as the number of cases of meningitis or diabetes in a population, but also to the relationship of that number to the size of the population. The resulting rate allows epidemiologists to compare disease occurrence across different populations.
>
> **Pattern** refers to the occurrence of health-related events by time, place, and person. Time patterns may be annual, seasonal, weekly, daily, hourly, weekday versus weekend, or any other breakdown of time that may influence disease or injury occurrence. Place patterns include geographic variation, urban/rural differences, and location of work sites or schools. Personal characteristics include demographic factors which may be related to risk of illness, injury, or disability such as age, sex, marital status, and socioeconomic status, as well as behaviors and environmental exposures.

Characterizing health events by time, place, and person are activities of **descriptive epidemiology**, discussed in more detail later in this lesson.

Determinants

Epidemiology is also used to search for **determinants**, which are the causes and other factors that influence the occurrence of disease and other health-related events. Epidemiologists assume that illness does not occur randomly in a population, but happens only when the right accumulation of risk factors or determinants exists in an individual. To search for these determinants, epidemiologists use analytic epidemiology or epidemiologic studies to provide the "Why" and "How" of such events. They assess whether groups with different rates of disease differ in their demographic characteristics, genetic or immunologic make-up, behaviors, environmental exposures, or other so-called potential risk factors. Ideally, the findings provide sufficient evidence to direct prompt and effective public health control and prevention measures.

Determinant: any factor, whether event, characteristic, or other definable entity, that brings about a change in a health condition or other defined characteristic.

Health-related states or events

Epidemiology was originally focused exclusively on epidemics of communicable diseases (*3* (http://www.cdc.gov/OPHSS/CSELS/DSEPD/SS1978/Lesson1/Section1.html#_ref3)) but was subsequently expanded to address endemic communicable diseases and non-communicable infectious diseases. By the middle of the 20th Century, additional epidemiologic methods had been developed and applied to chronic diseases, injuries, birth defects, maternal-child health, occupational health, and environmental health. Then epidemiologists began to look at behaviors related to health and well-being, such as amount of exercise and seat belt use. Now, with the recent explosion in molecular methods, epidemiologists can make important strides in examining genetic markers of disease risk. Indeed, the term health-related states or events may be seen as anything that affects the well-being of a population. Nonetheless, many epidemiologists still use the term "disease" as shorthand for the wide range of health-related states and events that are studied.

Specified populations

Although epidemiologists and direct health-care providers (clinicians) are both concerned with occurrence and control of disease, they differ greatly in how they view "the patient." The clinician is concerned about the health of an individual; the epidemiologist is concerned about the collective health of the people in a community or population. In other words, the clinician's "patient" is the individual; the epidemiologist's "patient" is the community. Therefore, the clinician and the epidemiologist have different responsibilities when faced with a person with illness. For example, when a patient with diarrheal disease presents, both are interested in establishing the correct diagnosis. However, while the clinician usually

focuses on treating and caring for the individual, the epidemiologist focuses on identifying the exposure or source that caused the illness; the number of other persons who may have been similarly exposed; the potential for further spread in the community; and interventions to prevent additional cases or recurrences.

Application

Epidemiology is not just "the study of" health in a population; it also involves applying the knowledge gained by the studies to community-based practice. Like the practice of medicine, the practice of epidemiology is both a science and an art. To make the proper diagnosis and prescribe appropriate treatment for a patient, the clinician combines medical (scientific) knowledge with experience, clinical judgment, and understanding of the patient. Similarly, the epidemiologist uses the scientific methods of descriptive and analytic epidemiology as well as experience, epidemiologic judgment, and understanding of local conditions in "diagnosing" the health of a community and proposing appropriate, practical, and acceptable public health interventions to control and prevent disease in the community.

Summary

Epidemiology is the study (scientific, systematic, data-driven) of the distribution (frequency, pattern) and determinants (causes, risk factors) of health-related states and events (not just diseases) in specified populations (patient is community, individuals viewed collectively), and the application of (since epidemiology is a discipline within public health) this study to the control of health problems.

Exercise 1.1

Below are three key terms taken from the definition of epidemiology, followed by a list of activities that an epidemiologist might perform. Match the term to the activity that best describes it. You should match only one term per activity.

A. Distribution
B. Determinants
C. Application

___ 1. Compare food histories between persons with *Staphylococcus* food poisoning and those without
___ 2. Compare frequency of brain cancer among anatomists with frequency in general population
___ 3. Mark on a map the residences of all children born with birth defects within 2 miles of a hazardous waste site
___ 4. Graph the number of cases of congenital syphilis by year for the country
___ 5. Recommend that close contacts of a child recently reported with meningococcal meningitis receive Rifampin
___ 6. Tabulate the frequency of clinical signs, symptoms, and laboratory findings among children with chickenpox in Cincinnati, Ohio

Check your answer.

References (This Section)

1. Last JM, editor. Dictionary of epidemiology. 4th ed. New York: Oxford University Press; 2001. p. 61.
2. Cates W. Epidemiology: Applying principles to clinical practice. Contemp Ob/Gyn 1982;20:147–61.

3. Greenwood M.Epidemics and crowd-diseases: an introduction to the study of epidemiology, Oxford University Press; 1935.

Previous Page Next Page: Historical Evolution of Epidemiology
Lesson 1 Overview

Lesson 1

Major Sections

Page last reviewed: May 18, 2012
Page last updated: May 18, 2012
Content source: Centers for Disease Control and Prevention (/index.htm)
Office of Public Health Scientific Services (/ophss/index.html)
Center for Surveillance, Epidemiology, and Laboratory Services (/ophss/csels/index.html)
Division of Scientific Education and Professional Development (/ophss/csels/dsepd/index.html)